DOCUMENT RESUME

ED 021 513 FL 000 983

By- Carroll, John B.; And Others
AN INVESTIGATION OF "CLOZE" ITEMS IN THE MEASUREMENT OF ACHIEVEMENT IN FORIEGN LANGUAGES.
Harvard Univ., Cambridge, Mass. Lab. for Research in Instruction.
Spons Agency- College Entrance Examination Board, New York, N.Y.
Pub Date Apr 59
Note- 142p.
EDRS Price MF-$0.75 HC-$5.76
Descriptors- *ACHIEVEMENT TESTS, BILINGUALISM, *CLOZE PROCEDURE, *EDUCATIONAL RESEARCH, ENGLISH, FRENCH, GERMAN, LANGUAGE ABILITY, LANGUAGE RESEARCH, *LANGUAGE TESTS, LISTENING COMPREHENSION, MEASUREMENT TECHNIQUES, RESEARCH METHODOLOGY, SECONDARY SCHOOL STUDENTS, SECOND LANGUAGES, STANDARDIZED TESTS, *TEST CONSTRUCTION, TESTS
Identifiers- College Board Achievement Tests

This study investigates the feasibility of using cloze procedure test items (in which a student supplies a word, letter, or phrase to fill a gap in a continuous text) for the written College Board foreign language achievement tests. An introduction which defines the problem, traces its history, and presents the overall design of the study is followed by a chapter on the development of cloze procedure test materials in English, French, and German. Other chapters discuss trial tests administered to English-French and English-German bilinguals, special studies of cloze test characteristics, the try-out of tests in secondary school language classes, and an experiment on the feasibility of auditory cloze procedure. A final chapter presents a summary, conclusions, and recommendations for further study, and appendixes contain test samples, questionnaires, and answer keys. (AF)

JOHN B. CARROLL

College Entrance Examination Board
Research and Development Reports

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

An Investigation of "Cloze" Items in the Measurement of Achievement in Foreign Languages

JOHN B. CARROLL, AARON S. CARTON, AND CLAUDIA P. WILDS

A Report on Research Conducted under a Grant from the College Entrance Examination Board, September 1957–February 1959

FL 000 983
ED021513

LABORATORY FOR RESEARCH IN INSTRUCTION
GRADUATE SCHOOL OF EDUCATION
HARVARD UNIVERSITY

Cambridge, Massachusetts
April, 1959

College Entrance Examination Board

RESEARCH AND DEVELOPMENT REPORTS

An Investigation of "Cloze" Items in the Measurement of Achievement in Foreign Languages

by

John B. Carroll, Aaron S. Carton, and Claudia P. Wilds

A Report on Research Conducted under a Grant from the
College Entrance Examination Board, September 1957--February 1959

Laboratory for Research in Instruction
Graduate School of Education
Harvard University
Cambridge, Massachusetts
April 1959

TABLE OF CONTENTS

Pages

Chapter 1. Introduction 1-12

The problem of the study, 1. History of the problem, 5. Overall design of the study, 10.

Chapter 2. Development of "Cloze Procedure" Test Materials in English, French, and German 13-26

Types of "cloze procedure" developed, 13. Selections of textual materials, 14. Preparation of word-cloze tests, 21. Letter-cloze tests, 24.

Chapter 3. Try-out of Written Tests with English-French and English-German Bilinguals 27-52

Subjects, 28. Experimental design, 30. Testing procedure, 31. Overview of results of word-cloze tests, 34. Reliabilties of scores on word-cloze passages, 39. Comparability of word-cloze tests across languages, 41. Results for the letter-cloze tests, 45. Summary of chapter, 52.

Chapter 4. Special Studies of Characteristics of Cloze Tests . 53-66

Procedure, 53. Scoring of tests, 57. The effect of paragraph cues on cloze scores, 58. Types of items aided by paragraph cues, 60. Comparative reliability of continuous and scrambled passages, 62. The nature of the ability required to do cloze items, 63.

Chapter 5. Try-out of Tests in Secondary School Foreign Language Classes 67-99

Cloze tests selected for the high-school study, 67. Test administration procedures, 68. Collateral data secured, 70. Scores on the cloze tests, 71. The level of foreign language proficiency of high-school students, 71. Intercorrelational results, 77. The reliability of cloze tests in the high-school groups, 81. Intercorrelations of word-cloze and letter-cloze tests, 84. Validity of the cloze tests as measures of foreign language proficiency, 84. Correlations of cloze tests with intelligence tests, 89. Correlations of cloze tests with foreign language aptitude tests, 90. Item analyses of word-cloze tests, 91. The use of community-of-response scores, 91.

Chapter 6. The Feasibility of Cloze Procedure in the Auditory Modality 100-106

Stimulus materials, 100. Test booklets, 101. Instruction to subjects, 101. Subjects, 102. Experimental design, 103. Treatment of data, 103. Statistical analysis, 103. Summary and conclusions, 106.

TABLE OF CONTENTS
(continued)

		Pages
Chapter 7.	Summary, Conclusions and Recommendations	107-119
References	. . . .	120-122
Appendix A	Adult Information Sheet	123
Appendix B	Sample Scrambled Test	124
Appendix C	Questionnaire and Cloze Instructions for French Word-Cloze Tests used in Secondary Schools	125-130
Appendix D	Questionnaire and Cloze Instructions for German Word-Cloze Tests used in Secondary Schools	131-136
Appendix E	Answer keys	137-138

Chapter 1

Introduction

The problem of the study

This is an investigation of the possibility of introducing certain relatively novel types of test material to supplement or replace the kinds of test items found in the written College Board achievement examinations in foreign languages. It should be said at the outset that for reasons of research strategy the main emphasis of this study has been upon the improvement of tests concerned with the written aspect of a foreign language. The importance of improving methods of testing competence with the spoken language is well recognized, but this study has devoted only minor attention to this problem because of the traditional concern of the College Entrance Examination Board with written tests.

The interested reader can find examples of typical CEEB foreign language test items in a pamphlet published from time to time by the Board, Foreign Languages: A Description of the College Board Tests in French, German, Latin, and Spanish. Objective test items of the sort found in this pamphlet characteristically yield highly satisfactory item-test correlations and are thus highly reliable, as has been pointed out by Paula Thibault (31). It seems to be the general opinion, also, that they are valid measurements of knowledge of a foreign language--at least of that kind of relatively intellectualized knowledge which is represented by the ability to state the meanings of foreign language words, the ability to choose correct grammatical forms, and the ability to answer questions based on printed prose paragraphs in the foreign language.

Use of these currently fashionable types of test items neverthless presents several difficulties, among which are the following:

(1) As in many other fields, the construction of good items is an expensive, time-consuming process requiring the services of subject-matter specialists carefully trained in techniques of test contruction.

(2) It has been argued that items tend to draw upon highly specific knowledges rather than upon a broad competence with the language as a whole. (This argument possibly overlooks the fact that the items nevertheless constitute a valid sample of language habits.)

(3) There is no satisfactory rationale for scaling the resulting test scores in terms of the extent to which native competence in the language is achieved.

(4) It is possible that there is a ceiling effect whereby conventional item types tend not to measure accurately at the upper levels of ability, because they measure the mere existence of language habits rather than their strength.

This study sought to investigate certain novel types of items which might circumvent some of the above difficulties and at the same time be characterized by satisfactory reliability and validity.

These "novel" types of items utilize what Wilson L. Taylor (26) has called the "cloze" procedure (after the "closure" which is supposed to occur when the subject supplies a word, letter, or phrase to fill a lacuna in a continuous text. They call upon the examinee's acquired stock of habits relative to what may be called the statistical contingencies of a language. Continuous texts in a language can be shown to exhibit many kinds of statistical regularity. For example, the probability that a noun will follow an adjective in certain contexts is very high. The probability that a vowel-letter will follow the letter sequence STR is well-nigh perfect. Whether and how much these regularities function in the composing of sentences as they are uttered or written is open to considerable question; nevertheless, the hearer of an utterance or the reader of a text, to the extent that he is well acquainted with the language, will develop expectancies as to what is going to follow any given part of the text. The more experience he has with the language, the better he will be able to predict how to complete a given context, or to fill in a lacuna. It seemed reasonable to suppose, therefore, that a valid method of testing competence with a language would be to ask the examinee to try to restore gaps or other types of mutilations in a text.

It was recognized that this method of testing language competence might have its own share of difficulties. Above all, a characteristic of the method is that it frequently requires "guessing," particularly in contexts where a number of different restorations of a text might be regarded as equally plausible or satisfactory. Under such circumstances it might be expected that even a person with a highly competent knowledge of a language might not be able to divine the "correct answer" (this being whatever stood in the original text). If guessing is frequent, reliability will inevitably decline.

Another difficulty is the very likely possibility that differences in "intelligence" or other extraneous dimensions of human variation might mask the variation in foreign language competence which this type of test seeks to measure. If such were found to be the case, it would be necessary to consider the advisability of trying to adjust for the influence of this disturbing variable.

Still another difficulty which was seen to be characteristic of this method of measuring language competence is the fact that the individual's ability to restore a text depends partly upon the nature and difficulty of the text itself. Indeed, Taylor developed the "cloze procedure" initially as a method for rating the difficulty or "readability" of texts (26); it was only at a later stage that he became interested in the possibility of using the technique for measuring individual differences in ability.

Despite these known technical difficulties, it was hoped that the cloze procedure in one or more of its possible variants might have certain advantages at least as a supplementary technique for the measurement of language competence. If the cloze procedure were found to be successful and if the technical difficulties could be obviated, it might have the following advantages:

(1) The technique offers a method of constructing test materials relatively cheaply and quickly. It would not call for so much time and attention on the part of subject-matter specialists as is the case with conventional test materials. One could readily construct materials for testing knowledge of any language provided appropriate texts were available.

(2) "Cloze" tests might measure aspects of language competence somewhat different from those measured by conventional tests, and it is conceivable that these competences might be regarded as more central to true mastery of the language.

(3) By using certain concepts available from Shannon's (24) information theory, it might be possible to justify a rational scale of language competence extending from complete absence of competence up to native or native-like competence.

(4) The tests might provide better differentiation at the upper levels of ability, discriminating between those with true mastery of a language and those with a more superficial knowledge of language patterns.

History of the problem

The "cloze" procedure was first invented by the German psychologist Ebbinghaus around 1897. Most graduate students in educational psychology are told that Ebbinghaus invented the ordinary "completion" or "fill-in" item which is so familiar in tests of intelligence and achievement. Actually, Ebbinghaus did no such thing; his kind of "completion" item was much closer to Taylor's cloze procedure. His test consisted of large quantities of Gulliver's Travels (in German translation) from which syllables had been more or less systematically deleted; the children were required to try to restore the original words. The test was intended to measure, not intelligence, but the degree of fatigue in mental functions experienced by school children at different times during the school day. Ebbinghaus found the test to be a poor measure of fatigue, but he noted that the test had high relationships with both age and excellence in school performance. (11)

The cloze technique also bears some resemblance to the method of controlled association developed first by Thumb and Marbe (see Woodworth [33, pp. 340-367] for an account of the history of the controlled association method) in the sense that it requires the subject to give a verbal response under certain contextual constraints. Howes and Osgood (15) have demonstrated certain properties of these contextual constraints in a controlled association experiment.

In 1939, Carroll (4) constructed a test of individual differences which also has some resemblance to the cloze item. This test, called "Phrase Completion", presented to the subject short incomplete phrases like

Hounds and_________
As for _________

and asked the subject to write the first response that came to mind. It was scored by community of response, that is, the response given to each item was weighted roughly in proportion to the frequency with which the response was given by a normative sample. After refinement through item analysis, the test was used in a factor analysis study conducted with a college-student sample; it was found to be one of the purest measures of 'V' (the verbal knowledge factor) available. It has since been used in several further factor studies (5, 10) and has consistently maintained its status as a pure measure of V. Its reliability, however, has not been satisfactory enough to recommend it for operational use. This is possibly because the test is too short and has not been subjected to any major effort to improve it. The reason for mentioning it here is to point out that a test requiring a subject to supply verbal material to fit a context--where the "correctness" of the response depends on the likelihood of the response in that context--has been found to be a good measure of the examinee's knowledge of his native language.

With the advent of "information theory" as developed chiefly by Shannon (24), other ways of asking subjects to respond to contextual constraints have been introduced and tried out. Much of the experimental work has been focussed on characteristics of texts and messages rather than on the differences in individual ability to respond. Shannon himself

(23) invented a special kind of guessing game for estimating the 'redundancy' of printed English; in this game, one attempts to guess each successive letter of a text either on the basis of the text already exposed or on the basis of the preceding n letters. Shannon estimated that printed English is about 75% redundant; that is, in theory one should be able to recode a text with about one-quarter of the number of characters and still convey the same information. Burton and Licklider (3) confirmed Shannon's estimate, using similar techniques. Chapanis (7) showed that random deletion of anything beyond 25% of the characters of a text made it exceedingly difficult to be restored to its original state, at least for most people. Miller and Friedman (20), however, have found that superior subjects, given practically unlimited time, can restore texts with up to approximately 50% deletions, and in view of certain other considerations this figure corresponds to a lower bound of about 60% redundancy. It seems to be well established, at any rate, that printed English has a redundancy of somewhere above 50%--certainly enough to make it feasible to construct tests requiring subjects to restore texts.

All the studies mentioned in the preceding paragraph were concerned with the guessing of letters deleted from a printed text. To the writer's knowledge, there are no studies of the guessing of phonemes deleted from a spoken text, although the studies of the intelligibility of speech with various kinds of interference as reviewed by Licklider and Miller (16) suggest clearly that speech has a high degree of redundancy. Miller and Selfridge (21) performed a "Shannon guessing game" with printed words in order to construct artificial texts with various degrees of statistical approximation to English.

In developing the "cloze technique", Wilson Taylor (26) has used tests which involve the guessing of deleted words, Taylor's chief purpose was to arrive at a measure of 'readability' of prose. The procedure calls for systematically deleting words in the sample of prose one desires to measure and submitting it to a panel of 25 or so people who are asked to guess the omitted words. The readability score of the prose passage is then computed as a function of the number of guesses which correspond perfectly to the omitted words in the original text. Readability scores determined by the cloze technique have been found to correlate well with those determined by other techniques, and in some cases they seem to correspond better to common sense assessment of readability. The cloze technique has also been shown to "work well" in Korean (29) and presumably it would work in other languages--that is, as a measure of the relative readability of prose passages. Cofer (8, 9) has used the cloze technique for slightly different purposes; in one study he used it to explore the characteristics of passages known to differ in adjective-verb quotient, and in another to index the trial-to-trial changes in the memorization of a prose passage.

It is convenient to adopt Taylor's term "cloze procedure" to apply to all types of test in which the subject is given a text with certain indicated deletions and asked to try to restore the original text. (Deletions are always indicated by replacing the deleted item--letter or word--with a blank of standard size.)

Individual differences in performance with the cloze procedure have been treated largely as a nuisance variable in many of the studies cited above. Shannon (23) dodged the problem of individual differences by using mainly one subject--his wife. Miller and Friedman (20), like many others, report the central tendency but not the variation of performance scores. Actually they are more concerned with the "best" or "maximal" degree of performance obtainable. As early as 1953, however, Taylor perceived the potentiality of the cloze technique as a way of measuring individual differences in language skill. His study, finally published in 1957 (30), showed that cloze scores attained by individual subjects were rather highly correlated with intelligence as measured by the Armed Forces Qualifying Test as well as with comprehension of, and success in learning, prose paragraphs concerned with certain technical subjects. Further test-retest reliabilities ranging from .74 to .88 were obtained by tests with 80 items given in a 40-minute time limit. Thus far, Taylor's study is the only published one of which the writer is aware that is explicitly concerned with cloze technique as a method of measuring individual differences.

The idea of using cloze technique (in one or more of its possible variations) as a method of measuring foreign language competence is a rather obvious development from the other uses to which cloze technique has been put. Taylor himself suggested (29) that it could be used in this way, but the first public suggestion to this effect seems to have been made by Victor Yngve of Massachusetts Institute of Technology in some informal remarks at the Northeast Conference on the Teaching of Modern Foreign Languages held at Brown University in the spring of 1954. Yngve's proposal suggested that the guessing of omitted letters in a foreign language text would be an effective technique. Bruner (2) seems to have been thinking along somewhat similar lines in speaking of an experiment which he and Robert Harcourt conducted at an international seminar at Salzburg. These experiments tested Italian, German, Swedish, French, Dutch, and English speakers on their ability to reproduce random strings of letters presented briefly, and third-order approximations to each of these languages. "As you would expect," writes Bruner, "there was no difference in ability to handle random strings, but a real difference in ability, favoring one's mother tongue, in reproducing nonsense in one's own language."

Overall design of the study

The present study is timely in the sense that it is one of the first to make investigations of the technical characteristics of cloze-technique items as measuring devices for individual differences. Our knowledge of the measurement characteristics of the cloze item is scant indeed; this study was designed to fill some of the gaps in our knowledge of the cloze-technique as a measure of competence in English, despite the fact that the main focus of interest was in the construction of foreign language tests.

The first step taken was to select texts from which test materials could be drawn. Since there was interest in comparing the effectiveness of the cloze procedure in various languages--English, French, and German--, it was thought necessary to assemble closely comparable materials in the three languages. Comparability was sought by locating materials existing in reasonably adequate translations in all three languages. From these materials a series of texts were drawn and made up into cloze tests in all three languages. Two types of cloze test were prepared: (1) the conventional cloze test as used by Wilson Taylor involving deleted words, and (2) sequences of letters of lengths 5, 7, and 11 from which letters were deleted either initially, medially, or terminally. This latter type of test was one of the types used by Miller and Friedman (20) in their studies of the statistical redundancy of English. The development of the tests is described in Chapter 2. For convenience, the tests will be designated word-cloze and letter-cloze, respectively.

An assumption underlying the measurement of foreign language competence is that the performance of adult native speakers of the language in question provides a criterion or standard against which the performance of the learner can be assessed. If one is truly measuring competence with a language as a general medium of communication, native speakers of a language are in some sense equal or uniform in their ability; linguists frequently seem to assume, at least, that native speakers are uniform in their knowledge of the language structure as the linguist defines it. Chapter 3 reports experiments designed to study the presumed degree of uniformity of language knowledge among native speakers. It also seeks to find out whether there are any differences in the redundancy of English, French, and German, using bilingual speakers as their own controls.

In the course of the experiments reported in Chapter 3, several disturbing methodological issues presented themselves. One, for example, had to do with the effect of context beyond the sentence. To what extent does the normal "logical" sequence of sentences in a paragraph supply contextual information over and above that contributed by the sentence itself? Another important question had to do with the extent to which cloze scores are affected by what may be called a "general ability to perform well on cloze tests," quite apart from knowledge of the language used in a particular cloze test. These and other problems are explored in Chapter 4.

Chapter 5 reports the results of a series of tryouts of some of the experimental tests in foreign language classes in several secondary schools, both public and private. Investigations of item validity, test reliability, and test validity are shown, as well as correlations with teachers' grades, intelligence tests, and College Board tests.

Chapter 6 reports a small exploration into the possibility of using cloze procedure with the spoken language. Since the procedure had scarcely been given a trial in any language, the experiment reported here was done on native speakers of English. One variable which was studied was the effect of manner of presentation--with and without "natural" sentence intonation.

As the reader may have noted in the brief review of the literature, no investigator has attempted to adapt cloze technique items to the format of machine-scorable objective test items. Nor did the present study attempt to see whether this was feasible for foreign language tests. Instead, it was conceived primarily as a brief pilot study to examine the general feasibility of the cloze technique, in several of its well-studied variants, as a method of measuring foreign language competence.

CHAPTER 2

Development of "Cloze Procedure" Test Materials in English, French, and German

Types of "cloze" procedure developed

If "cloze procedure" means any procedure in which one asks subjects to restore a mutilated text, one could legitimately employ almost any set of rules for mutilating a text. One would merely have to choose a unit for deletion and a rule for determining which units to delete. For example, in dealing with a printed text one could decide to delete selected parts of individual letters--say, the lower half of the printed line, as has been done in certain experiments on reading (1, pp. 194-195), or one might decide to delete the last inch of each printed line. It occurs to us to remark,incidentally, that "cloze procedure" is really nothing new for epigraphers and decipherers of ancient manuscripts, who have always had to contend with the 'deletions' caused by the ravages of time.

For the present study, it was decided to sample the possible range of deletion systems at two levels; the word, and the alphabetic character. In so doing, we were in effect following Taylor (26), who had used word deletion, and Miller and Friedman (20), who studied letter deletion. It seems possible that these two levels of deletion might call upon different kinds of language competence. Restoration of texts with deleted letters would require knowledge of the spelling of individual words and of the characteristic letter transitions of a written language system, while restoration of word deletions might depend rather on the ability to grasp the total meaning of written texts and on the availability of the individual's vocabulary in the language. Testing at both levels would, it was hoped, provide a well rounded picture of the individual's language competence as measured by the cloze procedure.

Selection of textual materials

The preparation of word-cloze tests (to be described in detail below) entailed the selection of a series of texts. There were two major considerations in the selection of texts: (1) texts were needed which existed in appropriate versions (the original, or translations) in all three languages for which it was desired to develop tests, and (2) the texts should cover an appropriate range of difficulty.

There were a number of reasons for deciding to select texts which existed in all three languages with which the project was concerned. The primary reason was that our objective was to establish difficulty equivalences across languages; that is, we desired to know whether, other things being equal, French and German were, respectively, easier or more difficult than English for native speakers doing cloze-procedure tests. There was also the necessity of appraising an individual's second-language competence relative to his competence in his native language, and such an appraisal, it was thought, could be best accomplished by having materials which could be established as being equivalent in difficulty across languages.

Covering an appropriate range of difficulty was important because we foresaw the need of selecting materials which would be of the optimal level of difficulty for appraising the foreign language competence of secondary-school students. Materials which proved to be too difficult even for discriminating among native speakers of a foreign language would surely prove to be too difficult for high-school students of that language. Difficulty, in this context, refers to the overall success or lack of success which subjects have in restoring a mutilated text. Taylor's work with "readability" has demonstrated wide variations in the difficulty of passages, and it is his thesis that "readability" or "ease of comprehension" can be measured as a function of the difficulty of a passage as obtained by the cloze technique.

Because of the considerable difficulty experienced in trying to locate text passages for which versions were available in English, French, and German, no attempt was made to sample the possible range of passage difficulty systematically. In any case, prior to the actual tryout of the passages there were no reliable measures of passage difficulty. To be sure, various readability formulas could have been applied to the passages in English, but they could not have been confidently applied to the passages in French or German. (Long after the passages had been selected, one of the original Flesch readability formulas was applied to the English passages, with results as indicated in Table 2.2. These results show that a wide range of difficulty was covered, even though not completely and not systematically.)

After a considerable amount of searching, it proved possible to locate texts which were available in English, French, and German version. Besides covering a reasonable range of difficulty levels, the searchers were able to secure for each language at least a few texts which were originally composed in that language. In order to avoid bias of results due to subjects' prior knowledge of the selections, it was considered important to select texts which would probably not be too widely familiar. The standard classics were generally avoided, while contemporary literature and the lesser known works of well-known authors were given prominence. Both fictional and non-fictional materials were sampled. Table 2.1 presents a list

TABLE 2.1

SOURCES OF TWENTY PASSAGES FOUND IN ENGLISH, FRENCH, AND GERMAN VERSIONS

The code designation of each passage consists of three parts: the symbol ç (for "cloze"), a number from 1 to 20, and E, F, or G for language. In several cases, two or more passages were drawn from each source; these are listed with corresponding page numbers for each edition.

Passage Number	Author	Reference
1, 2	Immanuel Kant	E.: Kant's Prolegomena, edited in in English by Paul Carus, Chicago, Open Court Publ. Co. , 1912 (ç1E, pp. 161-2; ç2E, pp. 82-3)
		F.: Prolegomenes, Paris, Libraire Hachette et Cie, 1891. (ç1F, pp261-2; ç2F, pp. 136-7)
		*G.: Prolegomena, heraus. v. J. H. v. Kirchmann, Berlin, Philos.-histor. Verlag, 1893. (ç1G, p. 150; ç2G, pp. 76-7)
3, 4	Alan Paton	*E.: Cry the Beloved Country, Chas. Scribner and Sons, New York, 1958. (ç3E, p. 238; ç4E, p. 130)
		F.: Pleure O Pays Bien Aimé, Traduit de l'anglais par Denise van Moppès, Albin Michel, Paris, 1950. (ç3F, pp. 374-75; ç4F, pp. 208-9)
		G.: Denn Sie Sollen Getrostet Werden, Fischer Bucherei, Frankfurt/M., 1951. (ç3G, pp215-6; ç4G, p. 117)
5	Margaret B. Johnstone	*E.: "The most valuable thing a man can spend," Reader's Digest, August, 1957. (Condensed from Guideposts) (ç5E, p. 82)
		F.: "Votre bien le plus precieux," Sélection du Reader's Digest, Oct., 1957. (ç5F, p. 40)

* Original version; i.e. the language of composition

TABLE 2.1 (continued)

Passage Number	Author	Reference
5	Margaret B. Johnstone (continued)	G.: "Unser kostbarstes Gut," Das best aus Reader's Digest, Oct., 1957. (ç5G, p. 97)
6	Max Eastman	*E.: "How human are animals?" Reader's Digest, Aug., 1957. (Condensed from Saturday Review) (ç6E, p. 115)
		F.: "Nos frères, les animaux," Sélection du Reader's Digest, Oct., 1957. (ç6F, p. 100)
		G.: "Wie menschlich sind doch Tiere!" Das beste aus Reader's Digest, Oct., 1957. (ç6G, pp. 59-60)
7	Wolfgang Langewiesche	*E.: "A new look at Niagara Falls." Reader's Digest, Aug., 1957. (ç7E, pp. 159-160)
		F.: "Un regard neuf sur les Chutes du Niagara," Sélection du Reader's Digest, Oct., 1957. (ç7F, pp. 45-6)
		G.: "Niafarafälle - geologisch gesehen," Das beste aus Reader's Digest. Oct., 1957. (ç7G, pp. 91-2)
8	Arthur C. Clark	*E.: "Secrets of the Sun," Reader's Digest, Aug., 1957. (Condensed from Holiday) (ç8E, p. 206)
		F.: "Les Secrets du Soleil," Sélection du Reader's Digest, Oct., 1957. (ç8F, pp. 1-2)
		G.: "Geheimnisvolle Sonne," Das beste aus Reader's Digest, Oct., 1957. (ç8G, pp. 41-2)
9, 10, 11	Georg Bernanos	E.: The Diary of a Country Priest, Translated from the French by Pamela Morris, McMillan Co., New York, 1956. (ç9E, pp. 22-3; ç10E, p. 237; ç11E, p. 46) *F.: Journal d'un Curé de Campagne, Plon, Paris, 1936. (ç9F, pp. 23-4; ç10F, pp. 202-3; ç 11F, pp. 43-4)

Table 2.1(continued)

Passage Number	Author	Reference
9 10 11	Georg Bernanos (continued)	G.: Tagebuch eines Landpfarrers, Fischer Bucherei, 1956. (ç9G, pp. 32-3) ç10G, pp. 292-3: ç11G, pp. 64-5)
12, 13	G. Révész	E.: The Origins and Pre-history of Language, Translated from the German by J. Butler, Philosophical Library, Inc., 1956. (ç12E, pp. 7-8; ç13E, p. 158)
		F.: Origine et Prehistoire du Langage, Traduction de L. Homberger, Payot, Paris, 1915. (ç12F, pp. 15-6; ç13F, p. 162)
		*G.: Ursprung und Vorgeschichte der Sprache, A. Franke A.G., Bern, 1946. (ç12G, pp. 18-9; ç13G, pp. 190-1)
14, 15, 16, 17	Marjorie K. Rawlings	*E.: The Yearling, Charles Scribner and Sons, New York, 1939. (ç14E, pp. 22-3; ç15E, pp. 398-9; ç16E, p. 200; ç17E, p. 348)
		F.: Jody et le Faune, Roman traduit de l'Americain par Denise Van Moppès, Albin Michel, Paris, 1946. (ç14F, p. 28; ç15F, p. 400; ç16F, pp. 204-5; ç17F, pp. 353-4)
		G.: Frühling des Lebens, Übersetzung von Maria Honeit, Rowohzt Taschenbuch, Hamburg, 1955. (14G, p. 22; ç15G, p. 370; ç16G, p. 184; ç17G, p. 324)
18, 19, 20	C. F. Ramuz	E.: When the Mountain Fell, Translated from the French by Sarah Scott, Pantheon Books, Inc., 1947. ç18E, pp. 28-9; ç19E, pp. 102-3; ç20E, p. 194)
		*F.: Derborence, B. Grasset, Paris, 1936. (ç18F, pp. 32-3; ç19F, pp. 118-9; ç20F, pp. 227-8)
		G.: Der Bergsturz, R. Piper and Co. München, 1936. (ç18G, pp. 24-5; ç19G, p. 90; ç20G, pp. 173-4)

TABLE 2.2

Passages arranged according to
the Flesch readability score
of the English version

Passage Number	Author	Flesch Readability Score **	Classification
17	Rawlings	.6	Reading Grade: 5.9 and below
14	Rawlings	.7	Description: Very easy. (difficulty of "light novel")
16	Rawlings	.8	
3	Paton	1.6	
20	Ramuz	1.9	Reading Grade: 6.0 to 6.9
15	Rawlings	2.0	Typical magazine: True Story
7	Langewiesche (RD)*	2.1	Description: Easy.
4	Paton	2.3	
11	Bernanos	2.3	
10	Bernanos	2.7	Reading Grade: 7.0 to 7.9
9	Bernanos	3.3	Typical Magazine; Liberty
			Description: Fairly easy.
19	Ramuz	3.9	Reading Grade: 8.0 to 9.9
8	Clark (RD)*	4.0	Typical Magazine: Reader's Digest
5	Johnstone (RD)*	4.1	Description: Average difficulty.
6	Eastman (RD)*	4.2	

* (RD) indicates selection from Reader's Digest

** The Flesch readability score is derived from the formula: (12)

$$\text{Readability score} = .1338X_s + .0645X_m - .0659X_h - .7502,$$

where X_s is a measure of words per sentence,
X_m is a measure of the prefixing and affixing of the words,
X_h is a measure of the number of personal, human interest, references.

TABLE 2.2
(continued)

Passage Number	Author	Flesch Readability score	Classification
12,	Révész	4.5	Reading Grade: 10 to 12.9 Typical magazine: Harper's Magazine Description: Fairly difficult.
18	Ramuz	5.2	
13	Révész	5.7	
1	Kant	6.9	Reading Grade: 17.0 and above (college graduate) Typical magazine: Scientific Monthly Description: Very difficult.
2	Kant	6.9	

of the sources finally chosen, with the exact page numbers on which the passages were found. Twenty passages in all, from 10 different authors, were made the basis of the word-cloze tests. Table 2.2 presents information showing the range of difficulty obtained for the passages in English.

Preparation of word-cloze tests

Two decisions have to be made in preparing a word-cloze test: (1) the length of continuous text from which deletions are to be made, and (2) the rule for deleting words within a stretch of text.

Most of Taylor's work with word-cloze has been concerned with the "readability" of texts over sizeable stretches, 175 words and up. He has thus allowed the subject to take advantage of the total force of context accumulating in rather lengthy passages. In preparing word-cloze tests in foreign languages, we expected many of our subjects to have competence far short of that possessed by native speakers, and we therefore desired to make the tests relatively 'easy'; one way of doing this was to use relatively lengthy passages, (205 words in length) from which subjects would be able to derive abundant contextual clues. Use of passages any longer than 205 words would have had the disadvantages of lengthening the total test beyond reasonable bounds, and incidentally of making some passages occupy more than a single double-spaced page.

The other decision which had to be made was that of choosing a rule for deletion. Taylor has tried various rules of deletion ranging from one word in five to one word in ten; for the measurement of prose readability he prefers to delete something like one word in every five or seven. Nevertheless, we believed that a lower rate of deletion might be more effective for measuring individual differences in language skill, desiring if anything to err on the side of giving the examinee abundant evidence on which to base his guesses. Further, we had no preconception as to what kind of deletion--in terms of 'parts of speech', word frequency, or other considerations--might be most effective; indeed, we wished to sample all kinds of word-deletions. In view of these considerations, it was decided to delete every tenth word. It should be stated here, however, that the present study has not attempted, and does not pretend, to provide a definitive answer regarding the optimal rate of deletion for measuring individual differences.

Criteria were established for counting words. In English and German, words were identified as strings of letters (including the apostrophe for contraction) preceded and followed by a space or mark of punctuation. In French, these criteria were slightly modified so as to count elided words like d' (as in d'argent) as separate words. Hyphenated words were always counted as two words, and occasionally a deletion fell on one member of a hyphenated word.

The 205-word passages always started at the beginning of a paragraph; naturally, they frequently broke off somewhere before the end of a complete sentence. Although the same passages were chosen in all three languages, it was hardly to be expected that they would cover the same amount of meaning-content. There did not seem to be any consistent parametric differences between languages, i.e. each language took about the same number of words to "say the same thing." Such variation as occurred could probably be attributed to idiosyncrasies of the writers and translators of the texts rather than to regular differences in language structure.

Each passage was mimeographed on a separate sheet. Counting from the first word, words 10, 20, 30, ..., 200 were deleted and replaced by a blank (using the "underline" character of a typewriter) of ten typewriter spaces. The blanks in each passage were numbered consecutively from 1 through 20.

Thus, all blanks except number 20 had nine words of unmutilated text both before and after it. In instances where a deletion would have fallen on a number, numerical expression, or proper name, the next word was deleted instead (without affecting the subsequent counting).[1]

Each version (English, French, and German) of each of the 20 passages[2] was subjected to the word-deletion procedure once; the resulting mutilated passages were used in the experiments in various combinations, as described in subsequent chapters.

The word-cloze tests were scored by counting the number of words for each passage which exactly corresponded to the words found in the original (completions which had even minor errors of spelling were counted wrong). According to Taylor (26), this scoring procedure is as effective as any. On a priori grounds, however, a case could be made for scoring in terms of community of response, i.e. giving a positive weight to any response of high frequency in a normative sample. Such a scoring procedure might enhance reliability and even validity because the "correct" response would correspond to a sort of linguistic norm rather than to the possible idiosyncratic item found in the original text. It should be remembered, also, that Taylor's recommendations are based largely on his work with the measurement of "readability" rather than individual differences in performance. To be sure, the 'readability' of a passage might quite well depend on the predictability of the words composing it, but the ability of readers to predict those words, depending rather on their acquaintance with the overall statistical characteristics of a language, might be better tested by means of a community-of-response scoring scheme. Some trials of such a scheme are reported in subsequent chapters. Nevertheless, it should be pointed out that the use of community-of-response scoring destroys one of the chief advantages which might be claimed for the cloze procedure, namely, that it enables one to dispense with extensive test construction and item analysis procedures.

---

[1] There are three instances in the corpus of test material where the deletion was inadvertently allowed to fall on proper names. In these cases the scoring admitted as correct either the exact proper name in the original, or an appropriate pronoun.

[2] The first 2 passages (from Kant) were omitted in the English-French experiment reported in Chapter III because of their extreme difficulty.

Letter-cloze tests

As in the case of word-cloze tests, comparability with the work of previous investigators was maintained also for the letter-cloze tests used in this experiment. In this case we replicated that procedure of Miller and Friedman (20) [actually Miller and Friedman studied several mutilation procedures] whereby strings of letters selected randomly from texts are presented with a blank indicating one letter missing, the subject being required to predict the missing letter. Miller and Friedman used strings 5, 7, and 11 characters in length and obtained data for each possible position of deletion. Ability to predict the missing character was found to be maximal for letters deleted towards the middle of the segment, falling off gradually as letters were deleted toward the beginning or toward the final position.

Because this study was concerned chiefly with the measurement of individual differences it was considered unnecessary to test Miller and

Friedman's method quite as exhaustively as they did. Still using strings of 5, 7, and 11 characters, we deleted letters only at the beginning, the middle, and the end of the string.

As in the word-deletion experiment, it was desired to compare the subject's performance in English and in the foreign language. There did not seem to be any good reason to draw the letter-deletion material from the texts used in the word-deletion tests, since it did not seem likely that the difficulty of letter-deletion material would be much affected by the overall difficulty level of the text. Instead, the English materials were drawn from those which had already been studied by Miller and Friedman. For the French and German, strings were drawn by a more or less random procedure from the October 1957 edition of the Reader's Digest in those languages. (Sélection du Reader's Digest and Das Beste aus Reader's Digest) The random procedure consisted of drawing a diagonal line down a page of text and copying the string intersected by this line in each successive line of text. Miller and Friedman excluded strings which contained proper names, numerals, and punctuation marks (admitting a space between a word, however, as a 27th "character" in the alphabet). The same procedure was followed for French and German, except that in French, the apostrophe was permitted as a character.

In each language, the test booklets contained 9 pages, each page containing 50 items of a given length (5, 7, or 11 characters) or place of deletion (initial, medial, or final). The order of pages for each language was 11a, 11b, 11c, 7a, 7b, 7c, 5a, 5b, 5c, where 'a', 'b', and 'c' denote, respectively, deletion of initial, medial, and final character. The English pages appeared first in each booklet but subjects were told they might do the pages in any order they pleased. At the top of each page were printed all the distinct permissible symbols for the language used there (including * for space between words). This meant that in French the subjects attention was drawn to the fact that such symbols as c, ç, e, è, é, â, à, etc. were to be regarded as different[1]; in German, subjects were shown a, ä, o, ö, u, ü as different symbols. Instructions and sample pages for the letter-deletion tests are to be found in the Appendix.

The scores for the letter-deletion test were the number of letters which exactly coincided with the letters deleted in the original text. The letter had to be supplied with the exact diacritical marking, if any. The usefulness of a scoring system based on community-of-response was also investigated, as reported in Chapter 5 . This was done despite the consideration stated earlier in this chapter that the use of item analysis procedures (necessary to establish community-of-response scoring) tends to vitiate some of the advantages claimed for the cloze procedure.

---

[1] Inadvertently, â was omitted from the French character list. This omission did not seem to inhibit anybody's using â when it was required. The â was never a correct answer. Nevertheless, it was used at least once. The error was rectified in the high-school testing.

CHAPTER 3

Try-out of Written Tests with English-French and English-German Bilinguals

In theory, or at least in a comfortably ideal world, competence in a foreign language should yield maximum and nearly uniform scores for native speakers of that language, because nearly every native speaker of a language acquires a certain set of language habits which are prerequisite to his functioning as an effective member of a speech community. The extent to which he goes beyond this minimal set of language habits is presumably a function of his general intelligence and his degree of education, but an ideal test of language competence should be unaffected by intelligence and degree of education.

Unfortunately, one can hardly expect to achieve this ideal. The very process of testing ordinarily calls upon certain cognitive skills which are independent of language competence and which, if present, will almost inevitably heighten test performance. If we can assume, however, that the individuals who are likely to take College Board foreign language examinations are relatively uniform in intelligence, degree of education, etc., --relatively, that is, in comparison with the total population--it makes sense to compare their performance with that of native speakers of the language who have somewhat similar degrees of intelligence and education. (We speak of "degree of education" not in the formal sense of number of years of education completed, but in the sense of general achievement.) This chapter reports some of the results obtained by applying cloze procedure tests to such samples of native speakers of French and German as we were able to obtain in the metropolitan Boston area. Because of the methods of recruitment employed, we found ourselves with data from a number of individuals whose native language was English but who also had high competence in French or German. Data from these subjects were found to be valuable as control data, and in any case, there was interest in seeing whether use of the cloze procedure would clearly distinguish between native speakers of a language and persons who have acquired a high degree of competence in the language as a second language. The bilinguality of the subjects was therefore a crucial factor in the design and analysis of this experiment.

Subjects

Subjects were obtained largely through the cooperation of associations of teachers of French and of German in the metropolitan Boston area. In each case, representatives of the project visited a regular meeting of the association and asked for volunteers to come to Harvard for what was estimated to be a two-hour testing session. Members attending the meeting were told that persons with a good knowledge of French or German were desired, with a preference for individuals with native facility. Emphasis was put on the necessity of establishing standards of performance on certain tests being considered for possible use in the College Board program.

Some individuals were recruited by a variety of other means--e.g. by inviting volunteers to ask their friends or relatives to participate, and by mail solicitation of persons listed as high school teachers of German. On the whole, it proved to be difficult to obtain any large number of subjects, and the persons actually obtained as volunteers were recruited only after rather considerable effort. (For example, the principal investigator went so far as to write and deliver a speech in German in order to comply with the requirement of the German teachers' association that no English be used at meetings.)

The total number of volunteers obtained in French was 17; in German, 22.

When the volunteers arrived for testing, they were first asked to fill out a questionnaire (See Appendix A) regarding native language, degree of education, and other relevant information. On the basis of their responses to a question regarding relative fluency in their two languages, subjects were classified as native English, native French, or native German. In one instance in which a subject reported that he was equally fluent in two languages, he was classified according to the country of his birth and youth. Several individuals were very difficult to classify because of very early bilingualism. At any rate, it can be safely said that every person listed as a native speaker of a language had spoken that language since early childhood, and on the record alone, many such individuals also had "native" or near-native proficiency in one or more other languages.

After all testing was completed, data were complete for the following numbers of cases, all of whom were adults (the numbers in parentheses are the number of cases who also completed the letter-cloze materials).

	English-French Bilinguals			English-German Bilinguals			Grand Total
	Men	Women	Total	Men	Women	Total	
Native German or Native French	2(1)	2(2)	4(3)	5(4)	7(4)	12(8)	16(10)
Native English	6(6)	6(6)	12(12)	9(3)	1(1)	10(4)	22(17)
Total	8(7)	8(8)	16(15)	14(7)	8(5)	22(12)	38(27)

All subjects, both native English and native French or German, had attained a relatively high level of formal education; they all had at least bachelor's degrees or the European equivalent. Several had master's or doctor's degrees or were in the process of acquiring them. The age range of the French subjects was approximately 22 to 60 years; of the German subjects, 24 to 60 years. Typically, subjects had acquired a knowledge of their second language by formal courses in it, although a few had acquired the second language (English) solely by absorbing it in the course of living and working in English-speaking countries. Most of the native English speakers had traveled or lived abroad in the countries of their second language.

In all cases, the subjects reported they were currently using both their languages in their daily lives or in their work. In this respect, the design of the experiment was fulfilled as well as might be expected; that is to say, the degree of bilingualism of these groups made it possible to make a meaningful comparison of performances in two languages, i.e. French vs. English, and German vs. English. (There were one or two trilinguals in the group, but it was not considered worthwhile to take advantage of this, particularly since the third language was reportedly not as well controlled by these individuals.)

Experimental design

For both the word-cloze and letter-cloze material, it was desired to compare subjects' performance in English and in the second language. In the case of letter-cloze material, it was possible to provide such an extensive sample of material in either language that there was no problem of insuring comparability of the basic material across languages, and it was not even considered necessary to use the same basic material (the same texts in different languages, that is to say) across languages. Therefore, all subjects were simply given all letter-cloze materials in both languages. The letter-cloze materials themselves varied in both length of string and position of deletion, as described in Chapter 2.

The case was somewhat different for the word-cloze materials. It was desired to limit the testing session to approximately two hours, and it was estimated that the subjects could not comfortably complete work on more than about 20 passages (each with 20 deletions from a text originally 205 words in length) in that time. Furthermore, it was obviously undesirable to give a subject the same passage in both languages. The experimental design therefore called upon each subject to do 10 passages in English and 10 other passages in the other language; subjects were paired, however, so that any passage which was done in English by one subject was done in the other language by the other subject. The assignment of passages in their English and foreign language versions was done by a randomizing procedure which varied from subject to subject, and the order of presentation of the passages in a test booklet was also randomized to help cancel systematic effects of practice and fatigue. Thus, each subject in the experiment had a uniquely-composed booklet. This design made it possible not only to assess the performance of the subjects in two languages but also to assess the comparative difficulty of the passages in the two languages and the comparative redundancy of the languages themselves.

Testing procedure

Each subject completed three distinct instruments:

(1) A questionnaire, mentioned previously, to yield data on how well the subjects knew the languages and how they acquired this knowledge;

(2) Twenty (for the German-English group) or 18 (in the English-French group) "word-cloze" passages, half in English and half in the second language;

(3) Eighteen sets of letter-cloze materials, of which nine were in English and nine were in the second language.

The first two of these instruments were filled out in group testing sessions at Harvard conducted by research assistants. Sessions for French and German were held at different times; because of the exigencies of scheduling it was necessary to give subjects the option of coming at one or another of several different sessions. During the intermission of each testing session, subjects were given instructions and materials for completing the letter-cloze materials; they were asked to do this work at home and return it by mail. (There was a loss of one French case and ten German cases for the letter-cloze experiment on this account.)

The detailed procedures followed in the testing will now be described.

In group testing, subjects were seated around a large table and given a brief explanation of the purpose of the experiment. They then filled out the questionnaire, after which they were told to read the instructions for the test. The text of these instructions follows:

> "In the passages that follow, every tenth word has been deleted. Please fill in each blank with the word that seems most probable to you. Do *not* try to be imaginative; put in the word that you think most people would put in that space. Only *one* word has been deleted each time, and all punctuation has been left in. Hyphenated words count as two words.
> We recommend the following procedure:
> 1. Read the whole passage through first without putting down any words.
> 2. Fill in all the blanks which seem fairly obvious and put down two or three alternative answers for the rest.
> 3. Select the most suitable alternative and reread the passage to be sure it makes sense.
> 4. Work quickly; do not spend too much time on any one passage."

(These instructions were original with this project. At the time they were prepared, Taylor's instructions were not available. Since then, comparison of these and Taylor's instructions shows only that our instructions lay more stress on trying to get the whole meaning of the paragraph before starting to fill in blanks.)

A few minutes were given for subjects to ask questions about the procedure. Subjects often asked whether they could go back to a passage after they had left it and gone on to another; they were told that they should go back to passages they wanted to work over only after they had finished the whole set of passages.

Before allowing the subjects to proceed with their test booklets, the examiner made the following additional remarks:

> "This is not a timed test. You are free to proceed at your own rate. You will find that some passages are harder and will take longer than others. We would, however, like to get an idea of how long each passage takes. Thus, if you will raise your hand and signal to the examiner when you have finished a passage and are ready to go on with the next passage, he will be able to jot your time down."

The examiner then said "Ready, begin!" and started his clock. As the subjects finished a passage, they signaled and went on to the next passage. The examiner kept a record of times for each subject.

In a few instances, subjects did not adhere to the timing instructions. They either neglected to raise their hands or they went through the test books doing the passages in one language first and then doing the passages in the other language. In these rare cases, the time record of the subject was not kept. In the first testing session it was feared that some subjects would fall far behind the others and take very long with the test. Thus after a rate for most subjects was ascertained--about 6 to 7 minutes per passage--the examiner once in a while suggested: 'You should be up to passage number ...' whenever he felt that one or another subject was falling too far behind. After the first session it was realized that subjects tended to finish booklets at about the same time and that some tended to start working quickly without accelerating their rates while others would start working slowly and get much faster as time went along. No further use was made of the times except to get a general idea of the time it took to do a passage. It may be said that the tests were administered under essentially work-limit conditions.

About half-way through the testing session there was a break for refreshments, during which the nature of the letter-cloze experiment was explained to the subjects. Informal questioning at the end of the test session revealed that the subjects found the test interesting and challenging but that they did not feel that they had done very well.

The entire testing session, including the introductory remarks and the break took about two and a half or three hours.

Overview of results of word-cloze tests

The main purpose of this experiment was to extablish normative data for the performance of native speakers of a language on the word-cloze tests. These normative data are presented in Tables 3.3 to 3.5, which contain the mean scores and other statistics for each passage as performed by one or the other of the several groups which were studied.

Passages 1 and 2 proved to be excessively difficult for the German speakers; consquently they were not used at all in the French group. Certain data on these passages will be presented, for completeness, but they will in general be excluded from statistical summaries.

For individuals doing passages in their native language, the mean passage scores (exclusive of passage 1 and 2 means) range in the following way: (Maximum possible score is 20 in each case.)

English	8.9	to	16.0	with a median at 11.9
French	9.5	to	17.0	with a median at 11.7
German	8.0	to	16.0	with a median at 11.7

For individuals doing passages in a second language the results are as follows:

French-speaking doing English:	6.0 to 15.0 with a median at 11.7
German-speaking doing English:	6.9 to 12.8 with a median at 9.1
English-speaking doing French:	7.3 to 17.2 with a median at 12.2
English-speaking doing German:	5.6 to 15.2 with a median at 8.5

It is evident that the passages vary considerably in difficulty, as might be expected. They also differ in their reliabilities, as indicated by their correlations with total score.

TABLE 3.1

Means, Standard Deviations and Correlations of Passages with Remaining Total Scores in English

English Passage No.	Native English Subjects				Native French Subjects[a]		Native German Subjects				All Subjects Combined			
	N	$\bar{X}$	s	r	N	$\bar{X}$	N	$\bar{X}$	s	r	N	$\bar{X}$	s	r
1	5	9.4	1.52	.33[b]	---	---	6	5.0	1.76	.87[b]	11	7.0	2.86	.87[b]
2	5	6.8	3.12	.75	---	---	6	4.6	2.15	.82	11	5.6	2.73	.74[b]
3	10	13.1	2.23	.40	2	15.0	7	11.1	2.12	.78	19	14.6	2.49	.56
4	11	14.9	1.51	.23	2	13.5	6	10.8	1.47	.91	19	13.8	2.44	.83
5	10	16.0	1.33	.46	2	13.0	7	11.4	2.51	.66	19	14.0	3.01	.80
6	12	10.3	1.50	.41	2	11.0	5	8.0	.71	-.78	19	9.8	1.86	.57
7	12	9.0	1.70	.45	2	10.5	5	7.0	1.22	.25	19	8.6	2.06	.56
8	10	13.9	1.29	.15	2	15.0	7	11.3	2.36	.56	19	13.1	2.27	.60
9	11	12.8	2.27	.52	2	10.5	6	9.2	2.14	.82	19	11.4	2.83	.62
10	11	9.7	1.62	.16	2	12.0	6	9.2	1.60	.36	19	9.8	1.76	.21
11	10	11.6	2.07	.53	2	12.5	7	6.9	2.73	.54	19	9.9	3.43	.64
12	11	15.4	2.34	-.04	2	14.0	6	12.8	1.94	.48	19	14.5	2.49	.41
13	14	10.6	1.01	.16	2	6.0	3	9.0	1.10	.00	19	9.8	1.82	.39
14	12	11.4	1.56	.34	2	10.5	5	10.0	1.87	.88	19	10.9	1.79	.60
15	10	12.2	1.48	.37	2	12.0	7	8.6	2.30	.58	19	10.8	2.66	.76
16	10	12.9	2.69	.70	2	11.5	7	7.9	2.34	.59	19	10.9	3.51	.96
17	13	13.9	1.54	.37	2	13.5	4	8.8	3.86	.73	19	12.8	3.14	.72
18	10	18.9	1.85	.62	2	19.0	7	8.4	.98	-.03	19	8.7	1.55	.28
19	10	10.7	2.11	.04	2	8.5	7	8.0	1.53	-.39	19	9.5	2.29	.44
20	11	11.0	1.94	.26	2	9.0	6	9.5	2.88	.32	19	10.3	2.33	.42

a- In view of the small N's only Means were computed for native French subjects.

b- For Passages 1 and 2 correlations of passage with total remaining score were computed on the basis of 20 passages.

c- For Passages 3 to 20 correlations of passage with total remaining score were computed on the basis of passages 3 thru 20.

TABLE 3.2

Means, Standard Deviations and Correlations of French Passages with Remaining Total Scores in French

Passage No.	Native English Subjects N = 6	Native French Subjects N = 2	All Subjects Combined N = 8		
	$\bar{X}$	$\bar{X}$	$\bar{X}$	s	r
3	14.2	15.5	15.0	1.51	.25
4	12.2	15.0	12.9	1.81	.36
5	12.7	14.0	13.0	1.85	.38
6	11.8	11.5	11.8	1.16	.17
7	9.0	10.0	9.2	2.25	.24
8	12.2	11.0	11.9	1.96	.63
9	7.3	10.5	8.1	2.03	.71
10	12.5	12.0	12.4	2.20	.65
11	13.8	10.5	11.8	1.67	.70
12	11.2	11.5	11.2	2.38	.56
13	13.0	14.0	13.2	1.49	.82
14	10.2	10.0	10.1	.99	.19
15	12.3	12.5	12.4	1.19	.43
16	14.2	13.0	13.9	1.81	.27
17	10.0	9.5	9.9	2.17	.69
18	8.3	11.0	9.0	2.51	.81
19	10.7	12.0	11.0	1.51	.53
20	17.2	17.0	17.1	1.36	-.16

TABLE 3.3

Means, Standard Deviations, and Correlations of German Passages with Remaining Total Scores in German

Passage No.	English Subjects N	$\bar{X}$	s	r	German Subjects N	$\bar{X}$	s	r	All Subjects Combined N	$\bar{X}$	s	r
1	5	4.8	3.11	.90	6	8.0	2.98	.90	11	6.5	2.98	.92
2	5	4.8	3.33	.63	6	4.3	1.86	.80	11	4.5	2.34	.53
3	6	15.2	1.94	.73	5	16.0	1.87	.12	11	15.5	1.86	.52
4	5	9.6	2.19	.86	6	11.8	1.60	-.30	11	10.8	2.14	.67
5	6	8.7	2.80	.82	5	14.4	2.70	.25	11	11.3	3.98	.80
6	4	9.2	1.71	.87	7	10.3	1.98	.75	11	9.9	1.87	.78
7	4	8.0	3.37	.84	7	9.9	2.34	.57	11	9.1	2.75	.70
8	6	8.3	3.27	.88	5	14.6	1.67	.86	11	11.2	4.14	.92
9	5	6.6	2.70	.82	6	8.7	2.25	.90	11	7.7	2.57	.87
10	5	5.6	2.30	.65	6	8.3	2.58	.64	11	7.1	2.74	.72
11	6	7.8	4.17	.95	5	11.6	1.82	.52	11	9.5	3.72	.91
12	5	6.6	2.74	.74	6	9.0	1.67	.80	11	7.9	2.16	.77
13	2	6.0	2.83	---	9	8.3	2.55	.76	11	7.9	2.62	.78
14	4	13.0	2.16	.95	7	14.6	1.27	.16	11	14.0	1.73	.71
15	6	9.8	2.14	.86	5	12.8	2.64	-.11	11	11.2	2.23	.82
16	6	9.3	1.97	.42	5	10.6	1.14	.71	11	9.9	1.70	.58
17	3	12.0	1.73	.40	8	13.6	2.88	.52	11	13.2	2.64	.53
18	6	6.8	3.37	.90	5	8.0	.67	.13	11	7.4	2.50	.76
19	6	7.7	4.03	.82	5	11.8	1.30	-.10	11	9.5	3.67	.83
20	5	11.4	2.30	.63	6	12.0	1.55	-.51	11	11.7	1.85	.13

TABLE 3.4

Summary Statistics on English Passage Scores

Statistic	Native English N = 22	Native French N = 4	Native German N = 12	All Subjects Combined N = 38
Mean of Passage Means	12.1	11.5	9.1	11.2
S. D. of Passage Means	2.08	----	1.57	1.77
Reliability of a single passage as a measure of individual differences (adjusted for passage differences)	.622	----	.245	.550
S. E. of measurement for a single passage	1.08	----	1.82	1.58

TABLE 3.5

Summary Statistics on French and German Passage Scores

Statistics	French Passage			German Passage		
	Native English N = 12	Native French N = 4	Combined N = 16	Native English N = 10	Native German N = 12	Combined N = 22
Mean of Passage Mean	11.8	* .	12.5	9.0	11.5	10.3
S. D. of Passage Mean	2.4	*	2.1	2.5	2.4	2.3
Reliability of a single passage as a measure of individual difference. (adjusted for passage differences)	.436	* ----	.365	.704	.549	.703
S. E. of measurement for a score on a single passage	1.29	* ----	1.46	1.51	1.27	1.52

* Not computed as N is small

Reliabilities of scores on word-cloze passages

The reliabilities of scores on each one of the cloze passages were estimated by the techniques propounded by Ebel (14). This technique is essentially an extension of intra-class correlation and uses the basic formulations of analysis of variance. It yields not only (1) the reliability of a single observation but also (2) the reliability of a composite score from k observations. In obtaining the reliability of scores on the cloze passages, we were faced with the difficulty that in general each subject had a score on a different set of cloze passages, because of the experimental design requirements. In Ebel's technique, one way of handling this situation is to ignore the variance due to passages, and this is justified if one is interested in the reliability of the total scores on a random set of k passages.

In computing reliabilities of single passage scores, it was desired to take account of the variance due to passage. This could not be computed in the ordinary way because of the difficulty mentioned above, viz., the fact that each subject had a score on a different set of passages, or stated differently, the fact that a given passage was not taken by all subjects. It was therefore, necessary to estimate the "sum of squares" for passages so that a proper error term could be determined. For the case of complete data it can be shown that the sum of squares for passages is equal to $N\sigma_{\bar{X}}^2$, where N is the total number of observations for all n persons and k passages, and $\sigma_{\bar{X}}^2$ is the variance of the means of passage scores. It was decided to use this relationship as a basis of estimating the mean square of passages. Thus $\sigma_{\bar{X}}^2$ was computed as the variance of 18 passage means. The sum of squares for passages ($S.S._p$) was then computed as $N\sigma_{\bar{X}}^2$. The sum of squares for error was then computed as:

$$(S.S._e) = (S.S._t) - (S.S._s) - (S.S._p)$$

where the three terms on the right are, respectively, the sum of squares for total, for subjects, and (as estimated) for passages. The number of degrees of freedom for error was $(N-1)-(n-1)-(k_o-1)$, where n is the number of persons and k_o is the average number of passages taken by each subject as computed by Ebel's formula (5):

$$k = \frac{1}{n-1}\left[\sum k - \frac{\sum k^2}{\sum k}\right]$$

The resulting mean square is taken as the variance of measurement for a score on a single cloze passage.

Pertinent reliability data are shown in Tables 3.4 and 3.5. The reliability coefficients themselves are a function of the range of variation of subject performance, but the standard errors of measurement are in theory independent of this variation. (In estimating the standard error of measurement for a test of double length as needed in Table 5.8, the values were multiplied by $\sqrt{2}$.)

It is of interest to note that the characteristic standard error of measurement of a single score on a cloze passage of 20 items is approximately 1.5.

Comparability of word-cloze tests across languages

It was hoped that the data collected in this experiment would yield information concerning the comparability of word-cloze scores in the three languages being studied. Serious problems arise in attempting to assess the data for this purpose. Ideally, large comparable random samples of native speakers should have been tested, comparability being established either by selection (common amount of education, for example) or statistically (by controlling on some relevant variable which would be identical for all groups, e.g. a non-language test of intelligence). Our data did not even begin to approach this ideal, and we did not have a proper control variable available.

Two techniques were employed to suggest provisional answers to the question. In the first of these, an analysis of variance was made of the mean total scores obtained by the three groups of native speakers--English, French, and German, each doing word-cloze materials in their own language. For each individual, the total score was obtained as the sum of whichever 9 passages out of passages 3 through 20 he happened to have worked on; if because of the randomization that had taken place, he happened to have done either 8 or 10 of these passages, the score was adjusted to a basis of 9 passages. Table 3.6 shows the means and standard deviations of the three score arrays, as well as the results of the analysis of variance. The F-ratio not being significant beyond the 5% level, it may be concluded that the word cloze materials in the three languages were of comparable difficulty to native speakers of those languages.

The second technique was a comparison by analysis of covariance of the second-language cloze-scores of native English and native German speakers, using their native-language cloze-scores as a control variable. (The number of native speakers of French was so small as to make it undesirable to apply

TABLE 3.6

Analysis of Variance
of
Total Scores of Three Native-Language Groups

Group	N	Mean	S. D.	Analysis of Variance Results
English	22	105.9	27.0	$S_b^2 = 143.8$
French	4	110.2	12.0	$S_w^2 = 510.6$
German	12	103.3	13.8	$F = \frac{143.8}{510.6} = .28$ Not significant
Total	38	104.9	22.15	

the technique to the comparison of the native English and the native French speakers.) The total scores used in this analysis were the total scores on passages 1 through 20; each total score was based on 10 passages. The results are shown in Table 3.7, and it is evident that there was a non-significant difference between the second-language cloze scores of native English and native German speakers. (Actually, this result shows little about the comparability of English and German word-cloze materials, although it indirectly assumes comparability; it is more relevant to the evaluation of whether the groups were equally bilingual, as they appear to have been.)

Our limited data suggest, then, that word-cloze scores on English, French and German passages are comparable, at least for native speakers of those languages.

A related question has to do with the consistency of passage difficulty across languages or across groups of individuals speaking different languages. These problems can be studied by obtaining rank-order correlations between various pairs of passage difficulty values reported in Table 3.1 to 3.3.

First, the consistency of the rank-ordering of the passages across groups was studied. The rank-order coefficients are in general high (all data are for 18 passages):

English passage difficulties	Rho
Between native English and native German speakers	.63
Between native English and native French speakers	.75
French passage difficulties	
Between native English and native French speakers	.74

TABLE 3.7

Analysis of Covariance of Second-Language Word-Cloze Scores (Y), Controlling on Native-Language Scores (X)

N = 10 native English + 12 native German = 22

	Total	Within	Between
Sum of Products	2785.2727	2947.3500	-162.0773
Sum of Squares for X	4467.0910	4020.3500	446.7400
Sum of Squares for Y	6859.3182	6799.7700	59.5510
Degrees of Freedom	21	20	1
Correlation Coefficient	.503	.564	---------
b_{yx}	.6235	.7331	
Adjusted Σy^2	5122.675n -	4960.020	162.655
Degrees of Freedom	20	19	1
Mean Square	256.134	261.054	162.655

$$F = \frac{162.655}{261.054} = .6231$$

F is not significant.

	Native Language	Second Language	
Means	$\bar{X}$	$\bar{Y}$	Adjusted $\bar{Y}$
German Subjects	108.334	89.083	92.037
English Subjects	117.300	85.800	82.181

German passage difficulties

	Rho
Between native English and native German speakers	.81

These results show simply that a passage in a given language tends to retain its relative order of difficulty regardless of whether it is applied to native speakers of the language or to persons who know the language only as a second language.

The more interesting question is whether the passages retain their relative difficulty values after translation into another language. The following results indicate that large changes may occur: (all data are for 18 passages)

	Rho
Between English passages performed by native English and the same passages in French, performed by native English	.19
Between English passages performed by native English and the same passages in French, performed by native French	.21
Between English passages performed by native English and the same passages in German, performed by native English	.33
Between English passages performed by native English and the same passages in German, performed by native Germans	.55

There is a suggestion here that despite the considerable changes that may occur in rank order, (a) consistency is greater between English and German than between English and French, and (b) consistency is greater when the passages are in every case performed by native speakers.

Results for the letter-cloze tests

The bilinguals who were tested with the word-cloze materials were given the letter-cloze booklets so that they might do them at home and return them. They were invited to call the experimenters at any time to ask questions about procedure. Although a few enjoyed the task, many found it onerous, and the final sample is not as large as might be desired, nor is it evenly distributed among native speakers of the three languages. Seven native Germans and five native English speakers took the English and German tests; three native French speakers and twelve native English speakers took the English and French tests. Thus there were totals of 27 subjects in English, 12 in German, and 15 in French.

For the score on each page in each language, and for the total score in each language, the following statistics are given in Table 3.8; mean, standard deviation, corrected split-half reliability (first 25 items vs. second 25 items), and standard error of measurement. The scores for each page represent the number (out of a possible 50) of correctly guessed letters, correctness meaning restoration of the actual letter standing in the original text from which the string was taken. The data in the table concern all cases available for any given test; thus, they include both native and acquired-language speakers of the language of the test.

The mean scores for English pages tend to be slightly lower than those obtained by Miller and Friedman, as we can show by converting our means to percentages and setting them against the percentages taken from Miller and Friedman's Figure 2 (19):

Percentage right
(letter-cloze tests)

Length of string and position deleted *	This experiment		Miller and Friedman
	All Cases (N = 27)	Only Native English (N = 17)	(N = 6)
5a	39	41	47
5b	52	55	61
5c	40	43	41
7a	52	54	56
7b	72	74	83
7c	54	57	64
11a	48	51	50
11b	90	92	94
11c	54	58	62

*a = initial; b = medial; c = final

TABLE 3.8

Results for Letter-Cloze Tests

Description of Sample	Language and Code Designation of Page*	Mean	S. D.	Reliability	Standard Error of Measurement
17 Native English; 3 Native French; 7 Native German 27 Total	English 5a	19.67	3.79	.588	2.43
	English 5b	26.07	5.47	.693	3.03
	English 5c	20.44	3.82	.612	2.38
	English 7a	25.85	3.31	.379	2.06
	English 7b	35.93	5.07	.850	1.96
	English 7c	27.00	4.50	.819	1.91
	English 11a	23.81	4.58	.724	2.40
	English 11b	45.22	3.78	.867	1.38
	English 11c	26.96	4.22	.737	2.17
	Total English	250.96	28.90		
12 Native English; 3 Native French 15 Total	French 5a	18.66	2.41	.163	2.20
	French 5b	30.93	3.46	.464	2.54
	French 5c	20.80	2.66	(-.316)**	2.66
	French 7a	25.13	3.25	.669	1.87
	French 7b	35.66	3.21	.530	2.20
	French 7c	24.86	2.84	(-.137)**	2.84
	French 11a	26.80	4.05	.628	2.47
	French 11b	42.47	3.57	.702	1.95
	French 11c	25.80	3.56	.698	1.96
	Total French	251.13	20.68		
5 Native English; 7 Native German 12 Total	German 5a	17.00	3.07	.482	2.18
	German 5b	23.33	5.22	.688	2.91
	German 5c	16.75	4.08	.250	3.54
	German 7a	20.42	4.67	.834	2.21
	German 7b	35.00	7.43	.904	2.96
	German 7c	21.25	4.12	.796	1.86
	German 11a	17.08	4.46	.481	3.21
	German 11b	41.67	5.84	.925	1.60
	German 11c	22.00	3.26	.576	2.12
	Total German	214.50	37.65		

---

* The code designation of the page shows the number of letters in the original text and the position of the deleted letter (a=initial; b=medial;c=final) For example, 5b means that the 3rd or middle letter of a sequence of 5 letters was deleted.

** Correlation between halves.

Our results for native English speakers are slightly closer to Miller and Friedman's, as we might expect. Miller and Friedman do not describe their small sample of subjects sufficiently well for us to judge comparability. The reader may be reminded that our English tests were precisely those used by Miller and Friedman.

Nevertheless, the general pattern of the results is similar to what Miller and Friedman found. Medial letters are much easier to guess than initial or final letters in a string, and the guesses for the longer strings are more likely to be correct because of the increased context.

What is remarkable is the considerable amount of variation in performance even among native speakers of the language of the test. Such variation, of course, is the basis for the rather high split-half reliabilities found for most parts of the test as well as for the total scores. The standard deviations of part scores are such as to suggest that the range of scores covers approximately from 30% to 70% of the total possible range, even for native speakers. The letter-cloze test seems to measure something besides pure language competence; this conclusion is reinforced by the finding that the correlations between total scores in German and English and in French and English proved to be .897 and .799 respectively. For bilinguals, performance in one's native language is highly related to performance in one's second language.

Some interest may attach to the relative ease of letter-cloze tests in the various languages. The t-tests between the means of total scores showed that for 15 English-French bilinguals English scores ($\bar{X} = 261.9$) were higher than French scores ($\bar{X} = 251.3$) at the 1% level of significance, and for 12 English-German bilinguals, English scores ($\bar{X} = 236.7$) were higher than German scores ($\bar{X} = 214.5$) at the .1% level of significance. While the mean difference of the English-French test may be due to the small number of native French speakers (only one of the three in the sample scored higher in French than in English), there was not a single instance where a native German speaker scored as high in German as in English. On the basis of information theory, a possible reason for the somewhat lower scores in French and German is the fact that they use an alphabet larger than the 27 characters (including * for space) used in English: in effect, French has a 35-character alphabet with all varieties of letters with accents and other diacritical marks, and German has a 30-character alphabet.

It was possible to make direct comparisons between French and German scores by an analysis of covariance which controlled the relative ability of the subjects on the basis of their English scores. The results are shown in Table 3.9. On the whole, there were no pervasive differences between French and German scores; the few cases where significant differences appeared seem to have been accidents of sampling or of test construction. Total scores, at any rate, showed no significant differences between the languages when controlled on English, either for the whole group or for the native English speakers alone.

Finally, correlations between letter-cloze and word-cloze total scores were computed for each group in each language. The results, shown in Table 3.10 indicate clearly that for both groups of bilinguals there are high correlations between letter-cloze scores in two languages, but much lower correlations between word-cloze scores in the two languages. For English-French bilinguals, word-cloze and letter-cloze scores tend to correlate substantially, particularly where the language is in common. For English-German bilinguals, a similar conclusion can be drawn, except for the fact that the German word-cloze test tends not to correlate significantly with any other score. The data, however, are based on an extremely small sample.

TABLE 3.9

Summary of Analyses of Covariance of French and German Letter-Cloze Scores Controlling on English Letter-Cloze Scores

(N = 15 English-French Bilinguals and 12 English-German Bilinguals)

Page	F-Ratio	Probability
5a	1.09	$P > .05$
5b	17.82	$P < .001$ (French better than German)
5c	7.70	$P > .05$
7a	4.46	$P > .05$
7b	3.69	$P > .05$
7c	6.14	$P > .05$
11a	21.23	$P < .001$ (French better than German)
11b	0.36	$P > .05$
11c	2.85	$P > .05$
Total	3.34	$P > .05$

TABLE 3.10

Correlations Between Total Word-Cloze and Total Letter-Cloze Scores

N = 15 English-French Bilinguals

		1	2	3	4
English Word-Cloze	1	1.00	.50	.61	.62
French Word-Cloze	2	.50	1.00	.36	.46
English Letter-Cloze	3	.61	.36	1.00	.80
French Letter-Cloze	4	.62	.46	.80	1.00
Mean		106.59	105.59	262.04	251.11
S. D.		10.25	9.64	18.91	20.68

N = 12 English-German Bilinguals

		1	2	3	4
English Word-Cloze	1	1.00	.06	.68	.65
German Word-Cloze	2	.06	1.00	.06	.14
English Letter-Cloze	3	.68	.06	1.00	.89
German Letter-Cloze	4	.65	.14	.89	1.00
Mean		95.3	93.7	236.7	214.5
S. D.		12.8	26.8	32.8	37.6

One hypothesis that should be further investigated is that letter-cloze tests put particular demands on the individual's spelling ability. A final word is in order concerning the reliabilities of the parts. The strings with the middle letter deleted tend to have distinctly higher reliabilities, and if letter-cloze tests are to be used for measuring second language proficiency it is advisable to construct them with seven or eleven character strings with the middle letter deleted.

Summary of this chapter

It is clear that native speakers of a language show considerable variation in their ability to restore texts in their native language when the texts are mutilated either by a word-deletion or by letter-deletion. Further, their ability to restore texts in a second language in which they have near-native proficiency, while slightly (and significantly) poorer than their ability to restore texts in their native language, is substantially correlated with the latter ability. This suggests that the ability to restore texts is somewhat independent of competence in a language as it is ordinarily defined. That is to say, we observe many people who are perfectly competent and literate in a language but who do not show facility in the special task of guessing what a missing letter or word in a text might be. If we wish to propose "cloze"-technique tests for measuring proficiency in a second language, it will be necessary to adjust for the individual's ability to perform "cloze" tests in his native language.

CHAPTER 4

Special Studies of Characteristics of Cloze Tests

The results of the tests of bilinguals raised a number of questions to which answers would be desirable if cloze tests were to be used in any foreign language achievement testing program. Among these questions were:

1. To what extent is the difficulty value of a passage sensitive to the particular set of words deleted?

2. To what extent is the performance of examinees on word-deletion materials dependent upon cues from the total passage, i.e. cues beyond the immediate context of a deleted word? Are these cues more potent when the paragraph is presented in its original form than when it is presented in scrambled form?

3. What kinds of items, from the point of view of syntactical structure, are most susceptible to the influence of paragraph cues?

4. To what extent is the same ability called for when paragraph cues are or are not provided?

5. What is the nature of the ability to guess deleted words? To what extent is it related to various established "factors" of cognitive ability?

Because of the lack of time, funds, and a plentiful supply of willing subjects, it was impossible to explore these questions as thoroughly as might eventually be desired. This chapter reports the results of a limited pilot experiment, as well as certain relevant results from a collateral study.

Procedure

Since the concern of this study was with the characteristics of cloze tests of native language ability, native speakers of English were used, and all materials were in English.

a. Test materials.

Seven passages (205 words in length) were selected from among the twenty which had been used in the study of bilinguals. They were selected as follows:

Two (4E and 16E) were to be used as the basis for control variables and had to have good reliability. From Table 3.1 it appears that the mean scores of all subjects on these passages were 13.8 and 10.9 respectively; they had reliabilities (correlations with total score) of .83 and .96, respectively.

Two (6E and 15E) had to be of medium difficulty and good reliability. Data from Table 3.1 show means of 9.8 and 10.8, and reliabilities of .57 and .76, respectively. They were to be used in studying the effect of varying the position of deletion.

Three (1E, 5E, and 12E) had to represent the range of difficulty used in the previous experiment. Data from Table 3.1 show means of 7.0, 14.0, and 14.5, respectively. Their reliabilities (.87, .80, and .41) were not considered particularly relevant to the problem under study, viz. the effect of paragraph cues upon the performance on individual items.

Passages 4E and 16E were administered to all subjects without modification.

For the experimental group, passages 6E and 15E were converted to what will be designated passages 6N and 15N, by restoring every deletion in 6E and 15E and then deleting the next word. Thus, while the original passages had the 10th, 20th, ...200th words deleted, the new passages had the 11th, 21st,....201st words deleted instead. (The usual constraints against deleting proper names and numerical expressions were observed.) The control group did the original passages 6E and 15E, however.

The experimental group also received "pied" passages which will be designated 1pi, 5pi, and 12pi. These were made up from passages 1E, 5E, and 12E by breaking them into 20 discrete items which retained the four words immediately preceding a deletion and the five words which immediately succeeded it; the 20 items were then placed on the page in random order, with the restriction that two adjacent sentences in the original would never be adjacent in the pied version. A sample page from a pied version is shown in Appendix B . It must have been obvious to most subjects that there were interconnections between the items in pied versions, but the instructions made no mention of this.

The control group received the original passages (1E, 5E, and 12E).

All subjects, both experimental and control groups, also received an additional page, designated Xpi, consisting of 20 items selected randomly from among the remaining 13 passages used in the bilingual study reported in the preceding chapter. As in the other pied versions, each item consisted of 10 words, the fifth of which was deleted and replaced by a standard-sized blank.

To make a preliminary study of the nature of ability to do cloze passage we administered to all subjects Thurstone's Four Letter Word test. This is one of the reference tests for the "Speed of Closure" (Cs) factor chosen by a committee of experts attending a conference on reference tests at the Educational Testing Service in 1951 (13). This test was selected for its brevity and presumed factorial purity; the speed of closure factor seemed to be the best one to represent the kind of "closure" which Taylor postulated to be the basis of his "cloze" technique. In Taylor's words,

"'Cloze' is derived from 'closure,' the term some psychologists use to refer to the notion that humans tend to perceive a familiar pattern as a whole even when parts of it are missing, obscured, or distorted." (28)

In the Four Letter Word test, subjects are required to locate familiar four-letter words embedded in sequences of otherwise random letters, e.g.

A M G E W I N D T E Y K Z C I R O C K W Q B H O W L O Z N P E B E L T O

where the words wind, rock, howl, and belt can be found.

It would have been desirable to try tests of other kinds of closure factors, but this was not feasible in this limited pilot experiment.

b. Experimental design and subjects.

Subjects were Harvard College upperclassmen and Harvard University graduate students, all paid for their services. Data from the original cloze experiment on bilinguals were included, when applicable, to supplement the data for the control group of this experiment.

Subjects were run either individually or in small groups. They were randomly (in order of appearance) assigned to the experimental group or the control group.

The experimental group (N = 20) received a booklet consisting of passages 4E, 16E, 6N, 15N, 1pi, 5pi, 12pi, and Xpi. Each subject's booklet had the pages in several different orders. Fourteen subjects of the experimental group were also given Thurstone's Four Letter Word Test.

The control group (N = 11) received a booklet consisting of passages 4E, 16E, 6E, 15E, 1E, 5E, 12E, and Xpi. That is, except for Xpi, all the passages were in their original form. Each subject's booklet had the pages in different random order, except that page Xpi was always last. Nine of these subjects were given the Four Letter Word Test.

The cover page of each subject's booklet presented the instructions for the cloze materials; these were virtually identical to the instructions which had been previously used in the bilingual experiment. Since the instructions referred to "the whole passage," one may assume that there was at least some tendency to consider even the scrambled passages as somehow integrated.

No time limit was set for any part of the test except the Four Letter Words test, which had a time limit of two and a half minutes.

Scoring of tests

Items were scored right only if the word inserted by the subject was exactly the same (apart from minor spelling errors) as the word which appeared in the text from which the passage was taken. Each page was scored separately; the maximum score possible on any page was 20.

The score on the Four Letter Word Test was the number of correct four letter words the subject succeeded in encircling in the allotted two and a half minutes.

The effect of altering the position of deletion

Taylor (27) studied two essays 175 words in length; these were mutilated by deleting every fifth word (35 in all) starting from five different initial locations to produce five different deletion versions. A significant overall difference was found in cloze scores attained by different fifths of a sample of 287 subjects. The particular words which are deleted in a passage therefore can be expected to make a difference in the overall cloze score for the passage. This is because individual deletions vary considerably in difficulty. In order to evaluate a passage accurately it is necessary to have a relatively large number of deletions. Taylor (27) has stated his opinion that values stabilize satisfactorily only when there are at least 50 deletions.

In the present experiment, it was possible to compare cloze scores from two deletion versions of each of two passages. Comparison was made by an analysis of covariance of the cloze scores from two groups (an experimental and a control group), holding ability to do cloze items constant by using scores from a third passage which both groups did in common. In one case, there was no significant difference between deletion versions, while in the other case, the difference produced an F-ratio which was significant at the .001 level. Our results tend therefore to confirm Taylor's; different deletion versions can and do sometimes produce significantly different scores for the passages. This conclusion holds for cloze passages in which there are 20 deletions occurring as every 10th word in a passage 205 words long.

The statistical comparisons on which the above statement is based are from (1) a comparison of scores on passages 6E and 6N holding cloze ability constant by means of scores on passage 16E and (2) a comparison of scores on passages 15E and 15N holding cloze ability constant by means of scores on passage 4E. In the first of these comparisons, data were available for 20 subjects in each group; in the second, there were 24 subjects in the control group and 20 in the experimental.

The effect of paragraph cues on cloze scores

These effects could be studied at two levels: (1) the effect of the paragraph in its original order vs. the effect of the paragraph when ten-word segments of it are scrambled or "pied" and (2) the effect of pied paragraph cues vs. the absence of such cues, as when ten-word cloze items are taken randomly from unrelated paragraphs.

The first level of effect was studied by means of three analyses of covariance. That is, scores on 1pi were compared with those on 1E, scores on 5pi with those on 5E, and scores on 12pi were compared with those on 12E. For the first two of these comparisons, scores on passage 4E were used as as controls; for the last, scores on passage 16E were used. It was necessary to use two different control tests because not all the adult bilingual subjects, who were used to augment the size of the control group, had taken both tests.

Significant ($P < .001$) effects of scrambling the items of a paragraph were found in all three instances. The adjusted means of the original and scrambled passage scores are as follows:

	Passage 1	Passage 5	Passage 12
Original	7.37	14.66	14.71
Scrambled	4.21	11.77	10.99
Difference	3.16	2.89	3.72

The design of the present study did not yield a precise test of the second kind of effect--that of having cues from the same paragraph, even when scrambled, as compared with having no cues. This would have to be done by an item-by-item comparison of items set in scrambled paragraphs with the same items assembled from unrelated materials, as in passage Xpi. Nevertheless, it was possible to compare all 60 available scores of the 20 subjects on the 3 scrambled passages (regarded as representative of the passages from which the items of the pied tests were drawn) with the scores of 30 subjects on the pied test Xpi. The former scores had $\bar{X} = 9.2$, $s = 3.8$; the latter had $\bar{X} = 7.6$, $s = 2.1$. A one-tailed t-test for uncorrelated means yielded $t = 2.3$, $P < .025$; this is a conservative test because some of the cases underlying the two means were identical and there is reason to expect considerable positive correlation between the scores. At any rate, cues from paragraph context seem to be influential even when the paragraphs are scrambled.

It was thought that paragraph organization cues might become more forceful toward the end of a cloze passage; if so, the mean scores would be higher toward the end of the passage. This seemed possible despite the fact that subjects were instructed to examine the entire paragraph before starting to work on it, because it was observed during testing that subjects tended to work from the beginning of the paragraph to the end, frequently without returning to the beginning or to earlier items.

This possibility was investigated by examining the difference in scores on the first 8 and the last 8 items in passages 1, 5, and 12 and comparing these differences under the non-scrambled and the scrambled conditions. According to the experimental hypothesis, the difference (score on the last 8 items minus the score on the first 8 items) should be larger under the non-scrambled condition (obtained in the control group) than under the scrambled condition (obtained in the experimental group). It was found, however, that the mean of the first 8 items was not regularly smaller than the mean of the last 8 items and that the differences between the mean differences never approached statistical significance. Thus, the notion that paragraph cues act cumulatively is not confirmed by these data. It is conceivable that paragraph cues sometimes work negatively, i.e. a paragraph cue could "throw off" the subject.

Types of items aided by paragraph cues

It was demonstrated above that context of paragraph length (as compared with a context of only 9 words) is a significant factor in permitting the subject to guess the word that stood in a particular position. Such a result agrees with theory: the more context, the more information, the better the guess. However, it is unlikely that context operates by the sheer weight of information; it is more likely that some words can be guessed through immediate contexts, other through more remote contexts.

Data from 22 control subjects (some from the bilingual experiment) and 20 experimental subjects were arranged so that the proportions of subjects getting an item correct when it was part of a continuous text could be compared with the proportion of subjects getting the same item correct when it was part of a scrambled text. Items showing a "marked" difference (20 or more percentage points) between the two conditions were identified. Of the 60 items in all, 26 showed a difference of this extent, but three of these were in the unanticipated direction, i.e. they were more often guessed correctly in the scrambled version. The words were classified according to conventional parts of speech. The following statements concerning the susceptibility of the conventional parts of speech to the influence of paragraph cues can be made only tentatively, since in most cases the samples are too limited:

Prepositions (including to when it occurs as a part of the infinitive) are highly stable. Only 3 out of 13 of these showed marked decrease in percentage under the scrambled condition.

Nouns show typically marked decrease in percentage guessed under scrambled conditions: 6 out of 13 showed this change.

Adjectives are relatively stable; only 1 out of 8 decreased 20% under scrambling.

Verbs are highly influenced by paragraph cues: 6 out of 8 showed 20% or more decrease in correct guesses under scrambling. The only verbs not showing this influence were the two auxiliary forms could and should.

There were too few examples of other form-classes to make any useful statements about them.

The results are consistent with the expectation that might well have been stated in advance: that "content-bearing" words such as nouns and verbs are more likely to be influenced by paragraph cues than words which are more concerned with the syntactical skeleton, so to speak, of the message. This result also suggests that when cloze technique involves connected discourse of paragraph length (as opposed to short sentence segments like the 10-word segments used here), the scores are more likely to be related to the examinee's ability to comprehend the total meaning of the paragraph, and thus perhaps to his general intelligence or verbal ability. We may hypothesize, therefore, that in order to achieve some suppression of the intellectual or verbal variance in cloze scores, paragraph-length materials should be avoided.

If this result had been discovered earlier in the project, more use would have been made, in subsequent experiments, of word-deletion tests utilizing short, 10-word segments.

The relative difficulty of items did not change radically as a result of scrambling. This was ascertained by determining the correlation coefficients between the array of difficulty values for the items in the connected passages and the array of difficulty values for the items in the scrambled passages. For pages 1E and 1pi, $r = .66$; for pages 5E and 5pi, $r = .74$; and for pages 12E and 12pi, $r = .68$.

Comparative reliability of continuous and scrambled passages

Ebel's technique for the reliability of multiple measurements (14) was applied to the scores of the 20 subjects on the 4 continuous passages and the 3 "scrambled" passages. The reliabilities were .43 for the total score on the 4 continuous passages and .72 for the total score on the three scrambled passages (or .77 when adjusted to be comparable with the figure given for 4 continuous passages). If continuous passages are truly less reliable than scrambled passages, as this result suggests, the explanation may be that the scores on continuous passages are influenced by the accidents of total meaning context.

The nature of the ability required to do cloze items

The results obtained with the Thurstone Four Letter Word test may be quickly summarized. There were no correlations between this test and any of the scores on cloze passages which were significantly different from zero. It appears, therefore, that the ability to do cloze items is not related to the speed of closure factor which is reportedly measured by the Four Letter Word test.

Fortunately, relevant data are available through the courtesy of F. D. Weinfeld . from a completely unrelated study (32) being conducted concurrently in the Laboratory for Research in Instruction. In this study, cloze passage 14 was included as a word-deletion test in a battery of 28 tests administered to groups of children in grade 9, 10, and 11. It was scored in the same manner as it was in the present study. Table 4.1 shows the correlations of this test with the 27 other tests in the battery, for 190 boys, and 154 girls, and for the total sample of 344 children. The tests are grouped on the basis of the factors disclosed in Weinfeld's study. Space does not permit a complete description of the tests; many of them come from the Educational Testing Service kit of reference tests. Those identified as being by Taylor are from Calvin Taylor's recent study of variables in communications situations (25). Instead of determining the average correlations of the cloze test with the tests loaded on each factor, indices of association were computed by treating the vector of correlations as if it were a column in the factor matrix in Weinfeld's study, normalizing all vectors, and then computing the inner products with the cloze test vector. The resulting indices are included in Table 4.1 .

It is immediately apparent that the cloze test scores were rather highly correlated with a variety of well-known cognitive factors. The cloze test correlates most highly with reasoning, verbal, theme writing, and

TABLE 4.1

Factor Loadings of Tests and Correlations of Cloze Test with Other Tests Used in Weinfeld's Battery (Grouped According to Factors), together with Indices of Association* (A) with Each Factor

Test Name and Number	Author	Factor Loading on its Factor	Correlation with Cloze Test Boys N =190	 Girls N=154	 Total N=344
Reasoning Factor: A = .760					
Sentence Order (10)	Adkins	.478	.523	.515	.537
Pedigrees (7)	Thurstone	.513	.496	.407	.489
Letter Series (4)	Thurstone	.518	.425	.310	.412
Reasoning (12)	Thurstone	.450	.376	.260	.309
Verbal Factor: A = .697					
Verbal Analogies (26)	Thurstone	.527	.553	.493	.541
Disarranged Sentences (27)	Thurstone	.575	.470	.412	.470
Completion (24)	Thurstone	.546	.524	.384	.468
Words in Sentences (25)	Carroll	.452	.422	.315	.403
Theme Writing (Themes rated on 4 scales): A = .591					
Sentence Structure (22)		.752	.509	.278	.439
Originality (20)		.699	.467	.284	.418
Choice of Words (21)		.800	.439	.274	.405
Organization (23)		.783	.451	.220	.394
Fluency of Expression: A = .540					
Inventive Opposites (5)	Thurstone	.339	.545	.523	.550
Word-Group Naming (14)	Adkins	.355	.505	.401	.470
Word Association (3)	Guilford	.249	.330	.241	.318
Telegram Writing II (13)	Taylor	-.331	-.293	-.175	-.279

---

* The indices of association, A_p, were obtained by computing, after normalization, the inner product of each vector in the factor matrix and the vector of correlation of the word-cloze scores with the tests in the factor matrix, that is, for the pth factor,

$$A_p = \frac{\sum_j a_{jp} r_{jc}}{\sqrt{\sum_j a_{jp}^2}\sqrt{\sum_j r_{jc}^2}}$$

where a_{jp} = the factor loading of test j on factor p, and
r_{jc} = the correlation of test j with the word-cloze score.

TABLE 4.1

(continued)

Test Name and Number	Author	Factor Loading on its Factor	Correlation with Cloze Test		
			Boys	Girls	Total
Word Fluency: A = .495					
Four-Letter Words (Production) (18)	Thurstone	.466	.385	.232	.348
Suffixes (16)	Thurstone	.370	.380	.243	.345
First and Last Letters (9)	Thurstone	.573	.271	.228	.272
Ideational Fluency: A = .400					
Distorted English (8)	Carroll	.296	.385	.412	.428
Letter-Star Test (2)	Carroll	.393	.333	.259	.331
Sentence Fluency (1)	Taylor	.472	.281	.137	.231
Multiple Completion Sentences (6)	Johnson	.443	.250	.142	.225
Topics (11)	Taylor	.625	.217	.144	.194
Plot Titles (19)	Guilford	.654	.210	.006	.152
Similies III (17)	Taylor	.722	.059	.052	.062
Thing Categories (15)	Taylor	.481	.057	.054	.041

expressive fluency factors as identified in Weinfeld's study. The cloze test is less associated with word fluency and ideational fluency factors, even though one might have expected a considerable association because of the apparent similarity of the tasks.

The fact that cloze scores are so highly correlated with various factors of cognitive ability when the testing is in the subject's native language raises grave question as to the potential efficacy of the cloze procedure as a measure of the subject's achievement in a foreign language.

CHAPTER 5

Try-out of Tests in Secondary-School Foreign Language Classes

The last major step of the present study was a tryout of the written cloze tests in a number of secondary schools, both public and private. We sought to find out something about the characteristics of such cloze tests as measures of foreign language proficiency, as well as to see how practicable these tests would be in a secondary school setting.

Cooperation was secured from 5 secondary schools in the region surrounding Boston; of these, 2 were public high schools and 3 were private schools. In every case, the groups contained at least a few students who had just taken, or were about to take, the College Board language examination in the language of their choice. Students of French were usually in their 3rd, 4th, or 5th year of instruction, but the students of German were all in either the 2nd or 3rd year of instruction. The testing was done in late April and early May, 1958; this was at a time when many of the students had just taken the College Board examinations.

Cloze tests selected for the high school study

The limitations of the testing time likely to be available made it necessary to use only a small selection of the cloze materials which had been developed and tried out with adults. Two passages were selected from the word-cloze materials and two tests from the letter-cloze materials. For French, these were passages 3 and 10 for the word-cloze and tests 11b and 7b for the letter-cloze tests. For German, passages 3 and 14 of the word-cloze materials, and pages 11b and 8b of the letter-cloze materials were used. These selections represented an attempt to offer high-school students highly reliable materials of not too high a level of difficulty.

Test administration procedures

In four of the five schools, the cloze tests were administered by classroom teachers according to written instructions which had been furnished to them; in the fifth school, the cloze tests were administered by project personnel.

The tests themselves consisted of 6-page mimeographed booklets. Page 1 contained a short questionnaire soliciting data on (1) how long the student had studied the foreign language; (2) whether he had studied any other foreign language; and (3) whether a foreign language was spoken in his home. (Except for the data received as to how long the language tested had been studied, little or no use has been found in this study for the information on this questionnaire. There were only sporadic cases in which foreign languages were spoken in the home.) (See Appendices C and D)

Also on page 1 were the instructions for the word-cloze tests which were to be found on pages 2 and 3. The instructions were as follows:

> Part I of this test is made up of two [French] or [German] passages, each from a different book. Every tenth word has been replaced by a blank. Your task is to read through the passage to see what it is about, and then try to fill in each blank with a [French] or [German] word that makes sense. The blanks are all the same length, but remember that the words which have been taken out may be short or long. As some blanks will be easier to fill than others, it is a good idea to do the easier ones first, and then go back to work on the harder ones, if you have time.
>
> Do one page at a time. When you finish one page, wait until your instructor tells you to go to the next. Do not look at the first page until your instructor tells you to begin.
>
> Work rapidly, because you have only nine minutes for each passage!

The nine-minute time limit for each word-cloze passage (20 completions) was chosen in the light of experience with adults. Even for the high-school students, it was long enough to permit most students to get to the end of the passage. It may be noted that the instructions asked the students to fill each blank with a word "that makes sense"; there was no suggestion that the student actually try to guess what word actually stood in the original text.

It is probable, however, that the precise wording of the instructions made little difference. The instructions were developed after fairly extensive tryouts with volunteer subjects obtained around the laboratory.

The two word-cloze tests followed as pages 2 and 3 of the booklets; they were separately timed.

Page 4 contained the instructions for Part II, the letter-cloze materials. The instructions for German, for example, were as follows:

> On the pages that follow are words and parts of words. The first page has examples 11 letters long; the second one has examples 7 letters long. An example does not necessarily begin at the beginning of a word or end at the end of a word, and the examples are completely unrelated to each other.
>
> Here the middle letter has been replaced by a blank; try to put a letter in the blank that makes sense. Besides the ordinary letters of the German alphabet, there is one other symbol you may use: "*", which stands for the space between words.

Sample Problems	Answers
JEDE*_OCHE*	JEDE*WOCHE*
ÄCHST_N*KAM	ÄCHSTEN*KAM
INMAL_SPOTT	INMAL*SPOTT
*ES_KLO	*ES*KLO
HM*_IT*	HM*MIT*
HEU_E*U	HEUTE*U

> Do not look at the next page until your instructor tells you to begin. When you finish one page, wait until you are told you may go on to the next.

Each page of the letter-cloze materials had listed across the top all the symbols (including variants with the several diacritical marks) which could be used. These arrays of symbols were intended to stimulate the student to indicate the exact letter and diacritical mark which he proposed to substitute for each blank. No special comment was made about diacritical marks, however.

The total cloze-material tests required nearly all of the typical class period. The schedule prescribed was as follows:

2 minutes: filling out information required on page one.
2 minutes: reading instructions to Part I and answering any questions.
9 minutes: first word-cloze passage.
9 minutes: second word-cloze passage.
2 minutes: reading instructions for Part II and answering any questions.
6 minutes: first letter-cloze page.
6 minutes: second letter-cloze page.

36 minutes (total)

Collateral data secured

(1) The Psi-Lambda Foreign Language Aptitude Test (Carroll and Sapon)

In some schools, it was possible to secure additional testing time in order to give the foreign language aptitude test developed by Carroll and Sapon (6). Use was made of the preliminary form identified as Form P, Psi-Lambda Foreign Language Aptitude Battery. Only parts 3, 4, and 5 of the test were given in order to conserve testing time. The authors supplied data permitting the establishment of T-scores[1] derived from a weighted combination of the scores on these parts. This test was given in order to afford a basis for matching the groups tested in terms of a variable related to success in language learning.

In all cases, this test was administered in person by representatives of the project.

(2) Results of CEEB language examinations.

Eighty-two of the 257 secondary school students tested in this experiment took the CEEB language examinations in French or German in the spring of 1958. Scores on these examinations were furnished either by the schools themselves or through the courtesy of the CEEB, New York. These scores utilize the College Board scale with mean of 500 and s. d. of 100.

---

[1] The T-scores had been scaled by the test authors in such a way that they had a mean of 50 and a standard deviation of 10 in a standardization sample of 912 cases, including about 11% high school students, 26% college students, and 63% adults (enlisted men in the U. S. Air Force, etc.)

(3) Teachers' grades.

These were the course grades given by the teacher for the term ending in June 1958; the scale in which they are given varies from school to school, some schools using letter grades, others using percentage figures.

(4) Academic average.

In some schools, data were obtained on the general academic standing of the pupils involved in the experiment.

(5) Intelligence measures.

A variety of measures of intelligence were available, all from school records. These included Scholastic Aptitude Test scores and IQ's derived from the Otis intelligence test.

Scores on the cloze tests

For the data summarized in this chapter, it may be assumed (except where otherwise noted) that the scores on the cloze tests were obtained by counting the number of completions which exactly corresponded to the words or letters which stood in the original texts. Both part and total scores were obtained for the materials.

The level of foreign language proficiency of high-school students

In dealing with any set of measurement data we are inclined to look first at measures of central tendency. In the present case we have a large number of small groups at various stages of their formal instruction in a foreign language. We may first ask: What level of proficiency do the tests show at various stages of instruction? Secondly: Do the tests differentiate among groups at various stages of instruction?

Table 5.1 shows the means and standard deviations of the several part

TABLE 5.1

Means and Standard Deviations of Cloze Tests Given to Secondary School Students (French), with Comparative Results from Adults

School	Year of Language Study	N	Word-Cloze						Letter-Cloze					
			Passage 3		Passage 10		Total		Test 7b		Test 11b		Total	
			$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ
A (Private Boy's School)	2 (Accel.)	10	11.9	1.3	6.7	1.3	18.6	2.6	25.9	6.3	25.7	7.1	52.6	11.6
	4	9	11.8	1.9	9.1	2.0	20.9	3.8	24.6	5.1	27.1	7.6	51.7	12.1
B (Private Girl's School)	3	12	9.8	2.9	4.6	2.4	14.3	4.9	21.2	4.5	22.3	4.5	43.6	8.3
	4	9	11.5	2.0	7.4	2.7	19.0	3.7	23.9	4.7	25.2	7.9	49.1	11.5
C (Private Boy's School)	3	26	8.9	2.3	5.7	2.4	14.6	3.6	19.4	6.0	20.8	7.0	40.2	11.5
	4	37	9.2	1.9	5.4	1.8	14.6	3.0	21.0	5.9	23.1	7.7	44.1	11.9
	5	23	9.3	2.9	6.9	2.8	16.1	5.1	23.5	6.1	26.0	7.0	49.6	10.6
D (Public High School) (Boys and Girls Combined)	3	38	9.0	2.1	5.5	2.1	14.5	3.6	23.7	6.5	22.5	8.0	46.2	13.0
	4	41	10.1	1.9	5.8	1.7	15.9	2.9	22.7	5.9	21.3	7.5	44.0	11.7
Adults: Native French	2 to 3		15.5	—	12.0	—	27.5	—	38.7	—	42.7	—	81.3	—
Native English	6 to 12		14.8	—	11.2	—	26.0	—	34.9	—	42.4	—	77.3	—
Combined	8 to 15		15.0	1.5	12.4	2.2	27.4	—	35.7	4.2	42.5	3.6	78.1	7.53

and total scores of the French cloze materials for a number of the separate groups. Comparative data from the adult testing described in Chapter 3 are also presented. Table 5.2 is a similar table of data for the German cloze tests.

In order to assist in the interpretation of the levels of proficiency indicated by the means, Tables 5.3 and 5.4 have been prepared. On the assumption that cloze scores measure on a ratio scale having an absolute zero, and on the further assumption that the performance of adult bilunguals represents an anchor point with which the performance of learning may be compared, the means for the various high school groups have been expressed as proportional parts of the means for adult bilinguals on the same tests. Because of the small number of cases available, the data for adult bilinguals are not as reliable as might be desired, although they are at about the same level in the two languages. In the case of the French data, however, there is an interesting proportionality between the results for the word-cloze and the letter-cloze tests. That is to say, the relative level of the mean word-cloze score of any group is very similar to the level of the corresponding mean for the letter-cloze test. For example, for all students in third-year French, the mean word-cloze score is 52.9% of adult performance, while the mean letter-cloze score is 56.0% of adult performance. Corresponding figures for students in fourth-year French are 58.7% and 58.0%, respectively. It is at least intuitively reasonable to think that the achievement of fourth-year students of French is about 60% of the asymptote of learning. On the other hand, the measured achievement of third-year students, at 52.9% and 56.0% of adult performance on the word-cloze and letter-cloze tests, respectively, seems higher than what one would expect. More information would be needed to interpret these results—e.g. information on the aptitude of the students and the nature of the instructional content.

TABLE 5.2

Means and Standard Deviations of Cloze Tests Given to Secondary School Students (German), with Comparative Results from Adults

School	Year of Language Study	N	Word-- Cloze						Letter-Cloze					
			Passage 3		Passage 14		Total		Test 7b		Test 11b		Total	
			$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ
C (Private Boy's School)	2	9	5.4	3.0	5.8	3.7	11.3	6.1	28.3	2.6	32.3	5.3	60.8	7.5
	3	5	7.0	.6	7.2	2.1	14.2	4.4	27.2	4.4	37.2	2.3	64.4	6.4
E (Public High School)	2	24	2.7	2.4	3.0	2.2	5.7	3.9	17.4	6.4	24.1	8.1	41.4	13.9
	3	25	4.2	2.9	5.2	2.9	9.4	5.4	20.4	6.1	23.7	7.2	44.4	11.5
Adults: Native German		5 to 7	16.0	1.9	14.6	1.3	30.3	---	33.3	8.8	41.1	6.7	74.4	15.3
Native English		4 to 6	15.2	1.9	13.0	2.2	28.6	---	37.4	3.7	42.4	4.2	79.8	7.6
Combined		11 to 12	15.5	1.9	14.0	1.7	29.5	---	35.0	7.4	41.7	5.9	76.7	13.0

TABLE 5.3

Means of French Cloze Tests for High-School Students Expressed as Proportional Parts of Means for Adult Bilinguals, By Year of Language Instruction

			French Word-Cloze Test		French Letter-Cloze Test	
Year	School	N	Mean	P.P.	Mean	P.P.
2	A*	10	18.6	.678	52.6	.674
3	B	12	14.3	.522	43.6	.558
	C	26	14.6	.533	40.2	.515
	D	38	14.5	.529	46.2	.591
	Total	76	14.5	.529	43.7	.560
4	A	9	20.9	.764	51.7	.662
	B	9	19.0	.694	49.1	.629
	C	37	14.6	.533	44.1	.565
	D	41	15.9	.580	44.0	.563
	Total	96	16.1	.587	45.2	.580
5	C	23	16.1	.587	49.6	.635
Adult Bilingual		Approx. 8	27.4	1.000	78.1	1.000

* This is an "accelerated" class.

TABLE 5.4

Means of German Cloze Tests for High-School Students, Expressed as Proportional Parts of Means for Adult Bilinguals, By Year of Language Instruction

Year	School	N	German Word-Cloze Test Mean	German Word-Cloze Test P.P.	German Letter-Cloze Test Mean	German Letter-Cloze Test P.P.
2	C	9	11.3	.383	60.8	.791
	E	24	5.7	.193	41.4	.539
	Total	33	7.2	.244	46.7	.607
3	C	5	14.2	.482	64.4	.837
	E	25	9.4	.318	44.4	.578
	Total	30	10.2	.349	47.8	.623
Adult Bilinguals		Approx. 12	29.5	1.000	76.7	1.000

The results for German (Table 5.4) do not show the clear proportionality which was evident for the French results. The means for the word-cloze tests are much lower than one would expect on the basis of the French results, and the means for the letter-cloze results are a little higher (relative to adult bilingual performance) than one would expect. These results may possibly indicate that different aspects of German are learned at different rates: those aspects relevant to a word-cloze test are learned relatively slowly, but those aspects relevant to a letter-cloze test are learned relatively fast. What these aspects may be is a matter for speculation and further research.

Both for the French and the German results displayed in Tables 5.1 and 5.2, it is disappointing that there is relatively little differentiation among groups in different years of language instruction. Even when we consider the year-groups in any individual school, there is not any dramatic progress shown from year to year. It has not even seemed worthwhile to make any precise tests of the statistical significance of the trends; many of the trends would undoubtedly be only of marginal significance, and even the trends which one would judge to be significant beyond the--say--5% level are still not strong enough to justify using the cloze tests as an indicator of amount of instruction.

Intercorrelational results

For the French cloze tests, the data were sufficiently voluminous in Schools C (a private boys' school) and D (a public high school enrolling both sexes) to justify the computation of intercorrelation matrices, shown in Tables 5.5 and 5.6. Table 5.7 shows intercorrelation matrices for School E (another public high school) where the German cloze tests were given. These data are the basis for the subsequent remarks concerning the reliability and validity of cloze tests.

TABLE 5.5

Intercorrelation Matrix for Cloze Tests, Language Aptitude Test, Grades in French, and CEEB French Score

School C (Private Boy's School)

N = 26 Year 3
= 37 Year 4
= 23 Year 5
N = 86 Total

					W-C		L-C					
			W-C T	L-C T					LAS	GIF	Mean	S.D.
		Yr.	1	2	3	4	5	6	7	8	9	
Word-Cloze Total	1	3	1.00	.30	.77	.80	.30	.24	.56	.63	14.6	3.6
		4	1.00	.19	.80	.79	.17	.17	.10	.42	14.6	3.0
		5	1.00	.52	.89	.89	.21	.61	.51	.72	16.1	5.1
		T	1.00	.35	.82	.84	.25	.36	.43	.57	15.0	3.9
Letter- Cloze Total	2	3	.30	1.00	.10	.37	.86	.90	.45	.43	40.2	11.5
		4	.19	1.00	.12	.19	.84	.91	.02	.16	44.1	11.9
		5	.52	1.00	.39	.55	.78	.83	.65	.37	49.6	10.6
		T	.35	1.00	.20	.38	.84	.90	.37	.37	44.4	12.0
Word-Cloze Test 3	3	3	.77	.10	1.00	.23	.14	.05	.48	.49	8.9	2.3
		4	.80	.12	1.00	.27	.18	.06	-.11	.06	9.2	1.9
		5	.89	.39	1.00	.59	.07	.53	.39	.64	9.3	2.9
		T	.82	.20	1.00	.38	.13	.21	.26	.35	9.1	2.3
Word-Cloze Test 10	4	3	.80	.37	.23	1.00	.33	.32	.40	.49	5.7	2.4
		4	.79	.19	.27	1.00	.09	.22	.27	.62	5.4	1.8
		5	.89	.55	.59	1.00	.31	.56	.53	.63	6.9	2.8
		T	.84	.38	.38	1.00	.27	.38	.45	.60	5.9	2.4
etter-Cloze Test 7b	5	3	.30	.86	.14	.33	1.00	.56	.58	.34	19.4	6.0
		4	.17	.84	.18	.09	1.00	.54	.06	.06	21.0	5.9
		5	.21.	.78	.07	.31	1.00	.30	.41	.24	23.5	6.1
		T	.25	.84	.13	.28	1.00	.51	.37	.28	21.2	6.2
etter-Cloze Test 11b	6	3	.24	.90	.05	.32	.56	1.00	.25	.42	20.8	7.0
		4	.17	.91	.06	.22	.54	1.00	-.02	.20	23.1	7.7
		5	.61	.83	.53	.56	.30	1.00	.63	.35	26.0	7.0
		T	.36	.90	.21	.38	.51	1.00	.28	.36	23.2	7.6
Lang. Aptitude Test	7	3	.56	.45	.48	.40	.58	.25	1.00	.67	54.5	7.2
		4	.10	.02	-.11	.27	.06	-.02	1.00	.25	52.8	5.2
		5	.51	.65	.39	.53	.41	.63	1.00	.57	58.4	6.4
		T	.43	.37	.26	.46	.37	.28	1.00	.58	54.8	6.6
Grade in French	8	3	.63	.43	.49	.49	.34	.42	.67	1.00	74.2	7.5
		4	.42	.16	.06	.62	.06	.20	.25	1.00	73.2	6.2
		5	.72	.37	.64	.63	.24	.35	.57	1.00	82.2	6.1
		T	.57	.37	.35	.60	.28	.36	.58	1.00	75.9	7.6
CEEB (N = 24)			.74	.34	.53	.79	.08	.46	.52	.80	541.8	89.0

TABLE 5.6

Intercorrelation Matrix of Cloze Tests, IQ, Grades in French and CEEB Language Scores

School D (Public High School)

N = 38 Year 3
= 41 Year 4
N = 79 Total

Variable		Yr.	W-C T	L-C T	W-C		L-C		IQ	GIF	Mean	S.D.
			1	2	3	4	5	6	7	8		
Word-Cloze Total	1	3	1.00	.48	.84	.84	.30	.54	.19	.47	14.5	3.6
		4	1.00	.39	.84	.80	.27	.39	.36	.33	15.9	2.9
		T	1.00	.41	.85	.82	.26	.44	.26	.40	15.3	3.3
Letter-Cloze Total	2	3	.48	1.00	.30	.51	.87	.92	.39	.18	46.2	13.0
		4	.39	1.00	.37	.28	.84	.90	.49	.16	44.0	11.7
		T	.41	1.00	.30	.40	.85	.91	.44	.17	45.1	12.4
Word-Cloze Test 3	3	3	.84	.30	1.00	.40	.11	.40	.20	.49	9.0	2.1
		4	.84	.37	1.00	.36	.24	.38	.38	.25	10.1	1.9
		T	.85	.30	1.00	.39	.14	.36	.28	.38	9.6	2.1
Word-Cloze Test 10	4	3	.84	.51	.40	1.00	.39	.50	.11	.29	5.5	2.1
		4	.80	.28	.36	1.00	.21	.27	.18	.31	5.8	1.7
		T	.82	.40	.39	1.00	.30	.39	.14	.30	5.7	1.9
Letter-Cloze Test 7b	5	3	.30	.87	.11	.39	1.00	.60	.31	.06	23.7	6.5
		4	.27	.84	.24	.21	1.00	.51	.45	.11	22.7	5.9
		T	.26	.85	.14	.30	1.00	.56	.38	.08	23.2	6.2
Letter-Cloze Test 11b	6	3	.54	.92	.40	.50	.60	1.00	.38	.25	22.5	8.0
		4	.39	.90	.38	.27	.51	1.00	.41	.16	21.3	7.5
		T	.44	.91	.36	.39	.56	1.00	.39	.20	21.9	7.8
IQ	7	3	.19	.39	.20	.11	.31	.38	1.00	.22	117.0	7.3
		4	.36	.49	.38	.18	.45	.41	1.00	.26	116.8	8.8
		T	.26	.44	.28	.14	.38	.39	1.00	.24	116.9	8.1
Grade in French	8	3	.47	.18	.49	.29	.06	.25	.22	1.00	52.9	21.1
		4	.33	.16	.25	.31	.11	.16	.26	1.00	54.9	20.9
		T	.40	.17	.38	.30	.08	.20	.24	1.00	53.9	21.0
CEEB (N = 25)		3	.59	.26	.41	.61	.17	.28	.24	.63	561.7	69.1
(N = 15)		4	.10	.38	.03	.16	.48	.18	.36	.37	581.5	58.7
(N = 40)		T	.43	.25	.30	.45	.24	.21	.27	.54	569.1	66.1

TABLE 5.7

Intercorrelation Matrix for Cloze Tests, Language Aptitude Test, IQ, and Grades in German

School E (Public High School)

N = 24, Year 2
= 25, Year 3

Variable		Yr.	Tot. W-C T 1	Tot. L-C T 2	W-C 3	W-C 4	L-C 5	L-C 6	LAS 7	IQ 8	GIG 9	$\bar{X}$	S.D.
Word-Cloze Total	1	2	1.00	.37	.88	.85	.27	.42	.55	.50	.79	5.7	3.9
		3	1.00	.70	.93	.93	.55	.67	.71	.61	.65	9.4	5.4
Letter-Cloze Total	2	2	.37	1.00	.29	.34	.93	.96	.08	.61	.11	41.4	13.9
		3	.70	1.00	.73	.58	.88	.92	.69	.48	.43	44.4	11.5
Word-Cloze Test 3	3	2	.88	.29	1.00	.49	.22	.34	.59	.59	.77	2.7	2.4
		3	.93	.73	1.00	.73	.58	.66	.67	.57	.53	4.2	2.9
Word-Cloze Test 14	4	2	.85	.34	.49	1.00	.24	.39	.34	.25	.59	3.0	2.2
		3	.93	.58	.73	1.00	.45	.59	.65	.56	.60	5.2	2.9
Letter-Cloze Test 7b	5	2	.27	.93	.22	.24	1.00	.78	.10	.52	.05	17.4	6.4
		3	.55	.88	.58	.45	1.00	.66	.56	.44	.23	20.4	6.1
Letter-Cloze Test 11b	6	2	.42	.96	.34	.39	.78	1.00	.07	.63	.15	24.1	8.1
		3	.67	.92	.66	.59	.66	1.00	.67	.38	.48	23.7	7.2
Language Aptitude Test	7	2	.55	.08	.59	.34	.10	.07	1.00	.50	.51	53.2	5.7
		3	.71	.69	.67	.65	.56	.67	1.00	.61	.71	54.4	7.1
IQ	8	2	.50	.61	.59	.25	.52	.63	.50	1.00	.43	114.2	7.8
		3	.61	.48	.57	.56	.44	.38	.61	1.00	.46	115.7	10.5
Grades in German	9	2	.79	.11	.77	.59	.05	.15	.51	.43	1.00	46.2	32.2
		3	.65	.43	.53	.68	.23	.48	.71	.46	1.00	48.4	26.5

The reliability of cloze tests in the high-school groups

It will be recalled that in Chapter 3 it was argued that an odd-even reliability computation would not be appropriate for cloze materials--at least not for word-cloze materials presented in continuous paragraphs--because the two halves would not be independent. Reliability was therefore computed on the basis of the interaction term (persons x tests adjusted for variance due to passages) in a situation where some eight to ten tests were available for each individual. For adult bilinguals, the standard error of measurement for a single cloze passage of 20 deletions was estimated as 1.46 (French) or 1.52 (German) and for a test of 50 letter-deletions as 1.37 (French) or 2.06 (German). Table 5.8 shows the standard errors of measurement for scores based on 40 deletions (word-cloze) and 100 deletions (letter-cloze).

For the high school samples, we have only the correlations between two tests; these tests are not strictly parallel because they are of differing difficulties. The intercorrelations between two non-parallel tests probably represent underestimates of reliability and, correspondingly, overestimates of the standard errors of measurement. Since no further data are available, we are forced to present these as the best available estimates of reliability. Table 5.8 shows these data for both French and German cloze tests, with estimated comparable data for adult bilinguals.

The reliabilities estimated for the French word-cloze test of 40 items range from .37 to .74 in various samples, to a considerable extent depending upon the variance of scores in these samples. Values of .55 and .56 are obtained for the School C and School D samples, respectively, with corresponding standard errors of measurement of 2.6 and 2.2. On the basis of these estimates, one may further state that a test of about 180 items would be

TABLE 5.8

Estimated Reliabilities and Standard Errors of Measurement of Total Word-Cloze and Letter-Cloze Scores, together with Data on the Intercorrelation of these Two Tests

			Word-Cloze Total (W) 40 Items				Letter-Cloze Total (L) 100 Items					Corr for attenuation
	Yr	N	$r_{\frac{1}{2}\frac{1}{2}}$	est. rel.	σ_t	$\sigma_{meas.}$	$r_{\frac{1}{2}\frac{1}{2}}$	est. rel.	σ_t	$\sigma_{meas.}$	r_{WL}	r_{WL}
FRENCH												
School C	3	26	.23	.37	3.6	2.9	.56	.72	11.5	6.1	.30	.58
	4	37	.27	.42	3.0	2.3	.54	.70	11.9	6.5	.19	.35
	5	23	.59	.74	5.1	2.6	.30	.46	10.6	7.8	.52	.89
	Total	86	.38	.55	3.9	2.6	.51	.68	12.0	6.8	.35	.57
School D	3	38	.40	.57	3.6	2.4	.60	.75	13.0	6.5	.48	.73
	4	41	.36	.53	2.9	2.0	.51	.68	11.7	8.6	.39	.65
	Total	79	.39	.56	3.3	2.2	.56	.72	12.4	6.6	.41	.65
Adult Bilinguals (estimated)*			.37 (N=16)	.54	3.0	2.0	.88	.93 (N = 15)	7.5	1.9	---	---
GERMAN												
School E	2	24	.49	.66	3.9	2.3	.78	.88	13.9	4.8	.37	.49
	3	25	.73	.84	5.4	2.2	.66	.80	11.5	5.1	.70	.85
Adult Bilinguals (estimated)*			.70 (N = 22)	.83	5.2	2.1	.90 (N = 12)	.94	13.0	2.9	---	---

* Data from Chapter III

required to produce reliabilities of .85 in these samples, and students would have to be allowed at least 81 minutes to complete such a test, based on our experience with a 9-minute time-limit with a 20-item test.

The reliabilities for the German word-cloze test (also of 40 items) are somewhat higher than the French, being .66 and .84 for the two samples from School E, but the standard errors of measurement are about the same as for the French word-cloze tests--2.3 and 2.2 in the two samples, respectively, indicating that the relative magnitudes of the reliability coefficients are largely a function of different ranges of talent.

The French letter-cloze tests for which reliabilities are estimated consist of 100 items; their reliabilities range from .46 to .75, typical values being .68 for School C data and .72 for School D data, with corresponding standard errors of measurement of 6.8 and 6.6, respectively. Nevertheless, it would still be necessary to lengthen the test by about two and a half times to achieve reliability in the neighborhood of .85; that is, the test would have to be about 250 items in length, requiring about 30 minutes as a time-limit if we extrapolate from the time-limit of 6 minutes used for a 50-item test.

Reliability was satisfactory for the German letter-cloze test. Coefficients of .88 and .80 were obtained, with $\sigma_{meas.}$ of 4.8 and 5.1 respectively, in two high-school classes in School E.

From Table 5.8 it can be seen that for word-cloze the standard error of measurement found for the high-school samples are comparable to the standard error of measurements estimated for adult bilinguals. This is true both for French and German word-cloze materials. It may be concluded that the word-cloze test has a relatively uniform standard error of measurement for different parts of the score range.

For letter-cloze tests, however, the standard error of measurement is distinctly smaller at the upper levels of the score range reached with the adult bilinguals. This seems to be true, at least, if we can judge from the lower $\sigma_{meas.}$ found for both French and German bilinguals as compared with the $\sigma_{meas.}$ in high-school samples.

Intercorrelations of word-cloze and letter-cloze tests

Together with data on the reliabilities of the respective tests, Table 5.8 presents data concerning the correlation between the word-cloze and letter-cloze tests. These statistics are pertinent to the question of whether these tests measure the same kind of language competence. Since the raw coefficients are partly dependent on the reliabilities of the respective variables, correlations corrected for the effect of attenuation have also been presented. In making these corrections for attenuation, the estimates of reliability made in Table 5.8 were used; since it is probable that these reliability values are under-estimates, it is thereby probable that the estimates of non-attenuated correlations are over-estimates. The independent values of non-attenuated correlations presented in Table 5.8 range from .35 to .89 with a median of .65. This value is far from unity, and in view of the probability that it is even then an over-estimate of the non-attenuated correlation, we may conclude that the word-cloze and letter-cloze tests measure somewhat different aspects of language achievement.

Validity of the cloze tests as measures of foreign language proficiency

From the limited evidence collected in the present study it is impossible to give a satisfactory assessment of the validity of the word-cloze and letter-cloze tests as measures of foreign language proficiency. A completely satisfactory judgment on this question could be probably made only after extensive factor-analytic studies of a variety of measures of language proficiency. Some tentative answers and suggestions arise from the present data, however.

One type of evidence of validity, of course, is the finding that the cloze tests differentiate between learners and those who have more or less completely learned, i.e. native or bilingual speakers of the foreign language in question. In a previous section it has been suggested that cloze-test scores measure on a ratio scale, and it has been pointed out that at least for the French cloze tests, word-cloze and letter-cloze tests yield group mean scores which represent highly similar proportions of adult performance. Such results strongly suggest that the cloze tests are in fact measuring some important facet of foreign language proficiency--but this is much more true for groups than for individuals. That is to say, if we use group results to cancel out individual variations on all the extraneous factors which may contribute to the determination of cloze test scores, the group means reflect real differences in foreign language competence.

Another kind of evidence of validity is to be found in the correlations of cloze test scores with teachers' grades. It is a common myth among educational psychologists that teachers' grades are notoriously unreliable; but this does not seem to be true, necessarily, of teachers' grades in foreign language courses, which are frequently found to correlate highly enough with other variables to suggest that they are quite reliable. This seems reasonable enough when we consider that the student's performance, especially in courses emphasizing the audio-lingual aspects of foreign language teaching, is much more concrete and readily apparent to any observer. Tables 5.5, 5.6, and 5.7 contain numerous correlations of cloze test scores with teachers' grades. The results are rather surprisingly variable and it is difficult to make generalizations. Nevertheless, as a general observation we may state that for the data with cloze tests in French there is a tendency for the word-cloze test to correlate higher with teachers' grades than the letter-cloze test does. For all cases at School C, word-cloze total scores correlate .57 with grades, while letter-cloze scores correlate .37 with grades. At School D, the correlations with teachers' grades are .40 for the word-cloze test and only .17 for the letter-cloze test. In view of the generally lower reliability of the word-cloze tests as compared with the letter-cloze tests, one could assume that the word-cloze test has an even higher underlying correlation with teachers' grades than might otherwise appear.

Similar results are obtained for the cloze tests in German. At School E, the word-cloze test showed correlations of .79 and .65 with teachers' grades, while the corresponding correlations for the letter-cloze test were .11 and .43.

These results, incidentally, constitute further evidence that the word-cloze and letter-cloze tests measure somewhat different aspects of language performance.

From the standpoint of validating the cloze tests, and on the assumption that teachers' grades constitute a valid criterion, it may be tentatively suggested that the word-cloze type is more valid than the letter-cloze type. Word-cloze tests make greater demands on the ability of the learner to select appropriate words and grammatical forms to fit a context, whereas the letter-cloze test demands chiefly a sensitivity to the orthographic customs of a language system and a knowledge of the proper spelling of foreign language words.

Further evidence pertaining to the validity of the cloze tests lies in the correlations between these tests and the CEEB language examination scores which were available for some cases. Tables 5.5 and 5.6 present several rows of correlation coefficients of part and total scores on the French cloze tests with CEEB scores. Again there is a surprising degree of variation in the results, but the word-cloze test shows correlations as high as .74 in one sample and as low as .10 in another. These samples are very small, however. The letter-cloze test has generally low correlations with the CEEB scores, ranging from .26 to .38.

In an effort to obtain more reliable results, data were assembled for all 76 students in the total sample (including both public and private schools) for whom CEEB French scores were available. (CEEB German scores were available for only 6 students and were not included in the analysis.) The correlation matrix for CEEB French scores with the various cloze scores is to be found in Table 5.9.

Of the two types of cloze tests, the word-cloze test has a higher correlation with CEEB French score, this correlation being .616 and the correlation for the letter-cloze test being .354. If we use the intercorrelation between the two parts of each test as a basic for an estimate of their reliabilities, we obtain (by the customary Spearman-Brown formula) a reliability coefficient of .656 for the word-cloze score and .683 for the letter-cloze score. Using these reliability values and assuming perfect reliability for the CEEB score, we can estimate the validities which the two types of cloze test would have if they were perfectly reliable. We find the correlations corrected for attenuation to be .761 for the word-cloze, and .429 for the letter-cloze score. These values suggest that neither type of cloze test measures exactly the same kind of achievement as is measured by the CEEB language test.

This conclusion further reinforced by the fact that the multiple correlation for predicting CEEB score from the two cloze scores is only .619, just barely greater than the zero-order correlation between word-cloze score and CEEB. Thus, the letter-cloze score adds hardly anything to the prediction of CEEB score from the word-cloze score. Whatever the letter-cloze test is

TABLE 5.9

Intercorrelations of Cloze Scores and CEEB French Scores

N = 76 cases
(Public and private secondary school
students for whom CEEB scores were
available)

		1	2	3	4	5	6	7
Word Cloze Total	1	1.000	.494	.868	.858	.264	.568	.616
Letter-Cloze Total	2	.494	1.000	.349	.506	.843	.898	.345
Word-Cloze Test 3	3	.868	.349	1.000	.489	.158	.424	.412
Word-Cloze Test 17	4	.858	.506	.489	1.000	.299	.558	.655
Letter-Cloze Test 7b	5	.264	.843	.158	.299	1.000	.520	.168
Letter-Cloze Test 11b	6	.568	.898	.424	.558	.520	1.000	.425
CEEB Score	7	.616	.354	.412	.655	.168	.425	1.000
Mean		15.99	47.75	9.75	6.24	23.79	23.96	608.8
S. D.		4.32	12.67	2.54	2.46	6.65	7.98	85.0

measuring, it is not related to CEEB-tested language achievement beyond whatever is accounted for by the word-cloze test.

The question of whether cloze-tests are better measures of foreign language achievement than CEEB-type tests can only be answered by research which would assess their validities with reference to a more ultimate kind of criterion than was available in the present study. Nevertheless, from the evidence presented here, particularly the evidence (Chapter 4) as to the kinds of verbal tests with which the word-cloze test correlates most highly, there is a suggestion that cloze-type tests are inferior as measures of foreign language achievement because they involve too much extraneous variance.

Correlations of cloze tests with intelligence tests

Further insight into the nature of the abilities measured by the foreign language cloze tests is gained by an inspection of their correlations with intelligence tests or with "IQ" ratings as found in school records. Data are available from School D, where French cloze tests were given, and from School E, where German cloze tests were given; the statistics are to be found in Tables 5.6 and 5.7.

In School D, the correlations tend to be higher for the letter-cloze test than for the word-cloze test, the values for combined groups being .44 and .26 respectively. On the other hand, at School E, the correlations are approximately similar: values of .61 and .48 are obtained for the letter-cloze test as compared to values of .50 and .61 for the word-cloze test.

Intelligence tests, as measures of verbal ability, would be expected to correlate rather appreciably with achievement in a foreign language, and this expectation is confirmed by these data. Nevertheless, to judge from the results with French cloze tests, verbal intelligence is more demanded by the letter-cloze tests than by the word-cloze tests. One may speculate that successful performance on a letter-cloze test depends to a considerable extent upon generalized verbal habits which relate to orthography and which are transferred between related languages; it is probable that intelligence tests measure the presence of such habits. Word-cloze tests, on the other hand, depend much more on actual knowledge of the language in question, however acquired; the writer's (unpublished) studies of aptitude for language learning show that intelligence is not necessarily related to success in language learning.

Correlations of cloze tests with foreign language aptitude tests

For several samples, correlations of cloze and other tests with the Carroll-Sapon language aptitude test were available, as shown in Tables 5.5 and 5.7. Close inspection will show that the pattern of correlations rather closely follows that of the correlations with teachers' grades. This result is not unexpected: since the language aptitude test is designed to predict teachers' grades and since it does indeed correlate quite highly with teachers' grades in the present study, it might be expected to correlate with any variable highly correlated with teachers' grades, as the cloze-tests generally are. This result suggests that the correlation between cloze tests (particularly the word-cloze tests) and teachers' grades reflects a common underlying variance which is central to foreign language learning; that is, that whatever else they measure, the cloze tests measure the central core of language achievement rather than some special variety of foreign language competence. This conclusion is not inconsistent with the observation that the cloze tests may not measure foreign language achievement very well, nor with the observation that the cloze tests may measure in addition some special competence in performing tests of the cloze-type, as has been suggested in Chapter 4.

Item analyses of word-cloze tests

For all available secondary school cases in French and in German, item analyses were performed for the 40 word-cloze items given to each of the groups. Item analysis consisted of finding the proportion of the total sample that got each item correct, and the biserial correlation of the item responses with total score on 40 items. The results are shown in Tables 5.10 and 5.11, the items being arranged in a series of somewhat ad hoc grammatical categories.

In terms of the proportions correct, it is clear that the "content-bearing" form-classes such as nouns, adjectives, and verbs are much more difficult to guess from context, while certain "function words" like prepositions and relative pronouns are easy to guess. Nevertheless, there is no consistent pattern differentiating the various form classes with respect to item validity. High item validities are found for both content-bearing form-classes and for function words; likewise, low validities can be found for both of these broad categories. It is likely that the validity of a particular word-cloze item is a function of the degree to which it is suggested by the context, and we may even go so far as to suggest that the item validity can be taken as a measure of the strength of context.

These item analyses suggest that in constructing word-cloze tests, there is little advantage to be gained by selecting particular kinds of words to delete. The use of any randomly selected words would produce results of the same nature.

The use of community-of-response scores

Up to this point, all the data reported are for cloze scores computed in the manner recommended by Taylor: the score is the number of items supplied by the examinee which are exactly the same as those standing in the original

TABLE 5.10

Item Reliabilities (r_{bis}) and Proportions Correct, by Part of Speech

French Word-Cloze Passages

Items 1-20 are from passage ç F 3; Items 21 -40 are from passage ç F 10.
N = 208 high school students

Item No.	Correct Response	Prop. Correct	r_{bis}
NOUNS			
20	magistrat	.423	.56
22	descente	.010	.40
25	note	.053	.38
33	temps	.072	.22
37	couloir	.019	.59
ADJECTIVES			
1	blancs	.029	.65
21	petit	.096	.62
28	prodigieuse	.014	.06
39	droite	.457	.33
VERBS			
16	continuaient	.010	.16
19	entendit	.043	.44
27	croyais	.000	---
32	etais	.582	.46
PRONOUNS			
4	qui	.384	.69
10	quoi	.120	.41
23	nous	.418	.41
29	s'	.231	.63
36	m'	.308	.70

Item No.	Correct Response	Prop. Correct	r_{bis}
ARTICLES			
6	un	.202	.56
12	les	.803	.14
17	la	.755	.41
18	les	.832	.26
OTHER DETERMINERS			
2	quelque	.048	.34
7	quelque	.604	.52
31	même	.875	.45
PREPOSITIONS (and PREP. LINKED WITH ART.)			
5	des	.344	.47
8	dans	.885	.29
13	au	.602	.23
14	de	.640	.57
15	en	.851	.41
24	sans	.625	.46
26	de	.779	.36
30	du	.130	.52
34	de	.731	.50
38	à	.053	.28

TABLE 5.10

(continued)

Item No.	Correct Response	Prop. Correct	r_{bis}
OTHER			
3	pas	.851	.28
9	trois	.010	.40
11	que	.226	.58
35	n'	.303	.42
40	et	.264	.36

TABLE 5.11

Item Reliabilities (r_{bis}) and Proportions Correct, by Part of Speech, German Word-Cloze Passages

Items 1-20 are from passage ç G 3; Items 21-40 are from passage ç G 14.

N = 63 high school students.

Item No.	Correct Response	Prop. Correct	r_{bis}
NOUNS			
9	Stock	.476	.54
15	Schlachtordnung	.000	---
20	Häuptling	.127	.43
28	Osten	.048	1.04
30	April	.000	---
32	Mutter	.333	.60
40	Jody (ihn)	.000	---
VERBS			
2	nahm	.238	.53
12	verwirrt	.000	---
14	stecken	.238	.89
24	wusste	.333	.59
35	traf	.000	---
36	erhob	.000	---
PRONOUNS			
10	ihn	.190	.67
11	ihn	.127	.77
40	Jody (ihn)	.000	---
ADVERBS AND ADJECTIVES			
18	wieder	.032	.63
21	wie	.238	.35
29	schwarz	.000	---
31	spät	.175	.65
ARTICLES AND DETERMINERS			
5	der	.619	.60
23	das	.571	.56
34	die	.508	.70
37	dem	.318	.76
38	die	.429	.66
39	das	.397	.25
27	anderen	.095	.37
33	andere	.286	.65
CONJUNCTIONS			
13	und	.190	.82
16	und	.048	.84
17	und	.238	.70
19	und	.079	.85
22	und	.365	.68

TABLE 5.11

(continued)

Item No.	Correct Response	Prop. Correct	r_{bis}
PREPOSITIONS			
4	zu	.095	.30
7	in	.540	.13
8	auf	.064	.21
26	zu	.079	.06
MISCELLANEOUS			
6	etwas	.159	.50
25	oder	.460	.78
1	nicht	.064	.30
3	nicht	.508	.33

text. In Chapter 2, however, it was argued that cloze scores based on a community-of-response scoring scheme might be a better method of measuring individual differences. A community-of-response scheme gives credit to a response in proportion to the extent to which that response is also given by the other members of some representative sample of examinees. For the present study it was decided, for convenience, to compute a community-of-response score by counting as "correct" not only the item standing in the original text but also any other "reasonable" completion given by at least 25% of the sample.* For example, in word cloze passage ç F 3, item no. 1 had the context "Le chef ne voulait pas être surpassé par les ________," the correct answer to which was blancs. 65% of the sample gave autres, however, which was also allowed as an answer. The additional answers allowed for all the word-cloze passages are as follows:

ç F 3

1. autres
2. les, des
9. les
19. et
20. chef

ç F 10

1. grand, vieux
2. route
3. moi
7. pouvais, voulais
8. longue, grande
9. nous
18. d', par, avec

---

* In determining what was a "reasonable" commonly given response, the data obtained from adults was also referred to. If 25% of the adults gave a certain response, it was included among the new "correct" responses for students, irrespective of how many students gave this response. In all cases a response had to be spelled correctly and given in the correct grammatical form to be counted as correct in the community-of-response scoring.

To study the community-of-response score, it was decided to draw a random sample of 100 cases from the French secondary-school groups only. No study of the community-of-response was done for German word-cloze passages and there was no study of letter-cloze materials in either language. Community-of-response scores were then computed and studied for reliability and validity.

For the conventional scoring system the correlation between passages 3 and 10 was .3861 yielding a stepped-up reliability of .5571. For the community-of-response scoring system the correlation was .4591, yielding a stepped-up reliability of .6293. To test the significance of the difference between the two raw correlations (.3861 and .4591), the procedure recommended by Peters and Van Voorhis (26, p. 185, [formulas 107 and 108a]) was used. This necessitated setting up the matrix of intercorrelations among the four scores involved, shown in Table 5.12. The test yielded a critical ratio of 1.35, which does not allow the rejection of the null hypothesis that the two scores are equally reliable. The added cost and effort required in establishing a community-of-response scoring key does not seem worthwhile, to judge from these limited results.

Unfortunately, for the 100 cases for which community-of-response scores were computed, the only external measure available to judge the relative validity of the two scoring systems as a measure of individual differences was the Language Aptitude Test (actually available for only 98 cases), already known to correlate rather well with language achievement measures. The total scores on Passages 3 and 10 scored by the conventional method had a correlation of .4627 with the Language Aptitude Test; when the passages were scored by the community-of-response method the correlation was .2630. The difference between the correlations, as tested by the procedure specified by McNemar (17 , p. 148, formula for t), is .1997 significant at the 1%

TABLE 5.12

Matrix of Intercorrelations among
Conventional and Community-of-Response
Scores for French Word-Cloze Passages
N = 100 secondary school students

Score			1	2	3	4
Passage 3,	Conventional Score	1	1.000	.385	.930	.501
Passage 10,	" "	2	.385	1.000	.462	.923
Passage 3,	Community of Response	3	.930	.462	1.000	.456
Passage 10,	" " "	4	.501	.923	.456	1.000
	Mean		10.06	6.24	12.19	7.57
	S. D.		2.35	2.43	2.46	2.84

level, thus we are inclined to conclude that the conventional scoring method is a better measure of individual differences in language achievement.

CHAPTER 6

The Feasibility of Cloze Procedure in the Auditory Modality

The bulk of the work which has been reported here, for reasons set forth in Chapter 1, has concerned the use of cloze procedure in written tests Taylor (29) has shown, however, that just as the "readability" of printed materials can be evaluated by means of cloze procedure, so also the "comprehensibility" of auditory messages can be measured by an analogous procedure. The present study being concerned with individual differences, it was desirable to determine whether cloze procedure in the auditory modality is an effective measure of individual differences. This chapter reports an experiment designed not only to explore the feasibility of auditory cloze procedure but also to investigate the effect of varying rate and intonation in the presentation of auditory cloze materials.

The study utilized cloze materials solely in English, in order to facilitate obtaining suitable subjects.

Stimulus Materials

The English forms of two passages used in other phases of this project were selected. These were passages ç3E and ç4E used in the study reported in Chapter 3. In order to increase the number of test items every seventh word was deleted instead of every tenth, and consequently each passage was divided into twenty-eight items. Each item consisted of six words preceding a deletion and four words following the deletion. Items subsequent to Item Number 1 thus started with the last four words of the previous item. Proper names were included in the text but were never deleted. When a proper name was a seventh word the word following it was deleted.

The test items were then recorded on magnetic tape; they were voiced by the experimenter. Two forms were prepared. One form is designated as the "expressive" form and the other is designated as the "non-expressive" form. In each item of the "expressive" form the experimenter, after reading the item number, read the first six words with normal expressive intonation at his normal speech rate, then a standard noise (produced by pressing the multiplication set-up key of a Monroe desk calculating machine) was substituted for the deleted word and the remaining four words were read with normal intonation and normal rate. In the "non-expressive" form the experimenter said the item number and then proceeded to utter the first six words of the item at the rate of one word per second. The words were read in a staccato monotone. As before, the noise of a desk calculator was substituted for the deleted word and then the remaining four words of the item were read in the same way as the first six words had been read. In both the "expressive" and "non-expressive" forms a pause of ten to twelve seconds was allowed before the tape went on to the next item. The pause was used by subjects to write their answers.

Test Booklets

Each subject was presented with a test booklet which consisted of a "Practice" sheet, which had four blank answer spaces, and two test sheets, consisting of twenty eight blanks for each of the passages.

Instruction to Subjects

A tape with instructions for subjects, which included a practice session, was played at the beginning of each experimental session. The text of this tape follows:

"This is an experiment to see how well people can replace words that have been taken out of spoken material. You do this perhaps unconsciously every daywhen chance noises make inaudible some of the words that are spoken to you and you have to guess what was being said. We want to explore systematically how people go about doing this. In order to get a record of the guesses you make several samples of English text were broken down into items. Each item consisted of several words; then a rattling noise; and several more words. It will be your job to write down on your answer sheet the word that you believe was covered up by the rattling noise. Time will be allowed for you to write your answer down before the test goes on to the next test item. You will not be allowed to hear an item more than once, but, because we are using continuous passages of text, each new item will always include a few words of the end of the previous item. Let us practice this procedure a bit before we start the experimental tests. Use your answer sheet labeled "Practice". Write your name and get ready to write your answer opposite Item 1. - - Pause - -

"Ready! Item One: 'Twinkle, twinkle little - -(noise)- - how I wonder ' - - Pause - -

"During the pause you should have written the word you think should have followed the word 'little', as for instance, 'star'.

"Item two: 'How I wonder - - (noise)- -you are, Up above'- Pause -

"Have you written the word that should have followed 'what'?

"Item three: 'Are! Up above the - - (noise) - - so high, like' - - Pause - -

"Item four: 'So high, like a diamond - - (noise)- - the sky,'

"The answers were, of course, Number One: 'star'; Number Two: 'you'* Number Three: 'world'; Number Four: 'in.' Notice that we never have more than one word for an answer. This is an important part of the test. You are to write only one word answers.

"Now we are ready to go on to the experiment itself. The material will be unfamiliar, and, we hope, will be much more difficult."

Subjects

Twenty volunteers were recruited as subjects from among the personnel of the Laboratory for Research in Instruction and the Harvard Graduate School of Education. All subjects had had at least some college undergraduate training. Most of them were graduate students.

---

* "What" was incorrectly taped into this place. Subjects were told the correct answer by the experimenter personally.

Experimental design

A Latin square design was used and arranged so that ten subjects were administered Passage A in the "expressive" form and Passage B in the "non-expressive" form. Ten other subjects were administered Passage A in the "non-expressive" form and Passage B in the "expressive" form. Within these groups of ten subjects, subgroups of five subjects each, were arranged so that the passage administered first was either A "expressive", or A "non-expressive", or B "expressive", or B "non-expressive". Table 6.1 shows the schedule of experimental procedures for all the subjects. The structure of the design is also made apparent in the tabulation of results in Table 6.2.

Whenever feasible, five subjects were experimented upon at once. But groups were not always to be obtained, and on occasion one subject was experimented upon at a time. Generally groups of two or three were experimented upon at the same time.

The experimental procedure lasted about thirty minutes. Subjects were asked for their reactions after the session. The ensuing conversations lasted from ten minutes to a half hour.

Treatment of the data

In tabulating the scores of subjects, an answer was scored as correct if, and only if, the word written by the subject was exactly the same as the word used by the author (or translator) of the passage. This criterion applied to the grammatical form of the word, as well. The sum of all such items scored as correct in a passage is the subject's score for that passage.

Statistical analysis

(I) The effect of altering the speech pattern

Table 6.2 is a tabulation of the raw data. Since the experimental

TABLE 6.1

Schedule of Experimental Treatments

Subjects	First Passage	Form	Second Passage	Form
1 thru 5	A	Expressive	B	Non-expressive
6 thru 10	B	Expressive	A	Non-expressive
11 thru 15	B	Non-expressive	A	Expressive
16 thru 20	A	Non-expressive	B	Expressive

TABLE 6.2

Tabulation and Sums of the Raw Data Classified as to Passage, Position and Experimental Constitution

	Passage A				Passage B				
	Position 1		Position 2		Position 1		Position 2		
	Subject	Score	Subject	Score	Subject	Score	Subject	Score	
Expressive Condition	1	10	6	15	11	15	16	10	$\sum X_e = 224$
	2	3	7	13	12	14	17	11	
	3	11	8	14	13	13	18	14	
	4	8	9	12	14	9	19	9	
	5	11	10	11	15	11	20	10	
Non-expressive Condition	16	12	11	10	6	14	1	13	$\sum X_n = 223$
	17	9	12	16	7	13	2	10	
	18	13	13	12	8	13	3	13	
	19	7	14	3	9	11	4	10	
	20	8	15	10	10	14	5	12	
	$\sum A = 239$				$\sum B = 208$				

design obviates the necessity for taking into consideration in the statistical analysis passage difficulty or position effects, a simple "t" of the difference between correlated means can be applied to the sum of scores obtained from summing the rows of cells. There is no statistically significant difference between the means in question. Hence it cannot be concluded that experimental variation in the speech pattern affects the difficulty of auditory cloze.

The number of subjects who answered an item correctly was taken as a score for each item. The scores in each condition were correlated with each other for each passage. The correlation coefficient, "r", between the "expressive" and "non-expressive" forms for passage A is equal to .8185. The correlation for passage B is equal to .8590. It can thus be seen that the variation in the experimental conditions did not affect the pattern of difficulty of the test items to any appreciable extent.

(2) Feasibility of the test form

As no effect due to experimental treatment was found, it is possible to ignore the classification of the data according to experimental treatment and to use the scores of each passage as though they comprised a half-test in order to compute a reliability measure. The Spearman-Brown reliability measure yielded by this procedure is .8342. Examination of the frequency distribution yielded by the test suggests that the scores are normally distributed. The data thus indicate that the test is a reasonably reliable measure of some kind of individual differences. However, the subjective reports of the subjects collected after each test session indicated that, generally speaking, the test procedure is found to be tedious, confusing and irksome. Subjects seemed to show a slight preference for the "expressive" form.

Summary and Conclusions

Two English auditory cloze tests were prepared on magnetic tape. The tests consisted of 28 items, each made up of six words, a deletion (signified by a noise) and another four words. The consecutive items of each test constituted a passage. Each passage was prepared in an expressive and non-expressive form. A Latin square design was used so that each passage and each form was presented to all twenty English speaking subjects. It was found that although the auditory cloze procedure proved to be a feasible measure of individual differences, as indicated from the reliability yielded by the data and the frequency distribution, the test form was unpleasant for the subjects. No difference was found between the expressive and non-expressive test forms. This finding suggests that auditory cloze scores will remain stable over a wide range of differences in intonation and rate of speakers. It does not imply that intonation plays no role in comprehension for it is probable that intonation carries information on a band of communication not affected by cloze procedure.

CHAPTER 7

Summary, Conclusions, and Recommendations

The present investigation was undertaken in order to ascertain whether "cloze" item types, in which the subject is required to restore a missing element in a string of continuous text, would provide a useful, reliable, and valid technique for measuring proficiency in a second language. It was hoped that tests composed of this type of item would: (a) be cheaper and simpler to construct than conventional tests, (b) draw upon a broad and representative sample of language habits rather than on a few specific knowledges, (c) yield a rational scale for measuring competence in terms of the extent to which native performance in the language is achieved, and (d) measure accurately at the upper levels of proficiency, where there is possibly a ceiling effect in the conventional item types.

Certain difficulties were anticipated, however. It was expected that "guessing", which is inherent in the procedure, would tend to reduce reliability; that "intelligence" and other dimensions of cognitive ability would prove to be disturbing variables; and that the "readability" of the texts would be a source of variation. Nevertheless, it was reasoned from the history of the use of the same and similar techniques, as well as from recent developments in information theory, that a measure of the ability to produce linguistic material consonant with the contextual constraints of the stimulus material might well prove to be an index of linguistic proficiency.

The study was conducted as a pilot investigation in order to explore a number of facets of the problem.

Two types of items were investigated: word-cloze and letter-cloze. The word-cloze items were modeled very closely after the materials used by Wilson Taylor (26) and named "cloze procedure" by him. They consisted of passages of continuous text 205 words in length in which every 10th word was deleted. The letter-cloze items were an adaptation of test items used in an experimental study by Miller and Friedman (20) and consisted of strings of 5, 7, 9, or 11 letters (including spaces as letters) in which the first, middle, or last letter had been deleted. In both types of test, it was the task of the subject to attempt to restore the linguistic material which had been deleted.

Twenty word-cloze tests in English, French, and German were drawn from a corpus of ten prose selections for which English, French, and German versions were available. The basic materials ranged in complexity from Readers' Digest articles to Kant's philosophical writings. Application of one of Flesch's original readability formulas (12) showed that the sample was representative of all Flesch readability levels. English letter-cloze materials were supplied through the courtesy of Miller and Friedman, and French and German letter-cloze materials were developed by applying Miller and Friedman's method to appropriate texts. The materials were presented in typewritten, mimeographed form. Deletions were always indicated by the underline character of the typewriter.

A word, in English and German, was defined as a string of letters bounded by a space at the beginning and end. In French, allowances were made so that morphemes such as d' (e.g. in d'argent) could be counted as words. In most studies conducted here, responses were scored as correct only when the deleted word or letter was replaced exactly as it appeared in the original test. One study tested the feasibility of using a community-of-response criterion.

Groups of English-French and English-German adult bilinguals of relatively high academic achievement were tested with these materials in order to establish a norm of native-speaker performance and to compare results for the three languages. The mean scores of passages ranged rather widely and significantly as expected. Versions of the same material varied in their relative difficulty over the three languages. The individual passages varied, too, in their ability to predict the total score on all passages. Total reliabilities of word-cloze tests consisting of 9 passages in the same language ranged from .245 to .704. Scores on letter-cloze tests proved to be dependent upon the number of letters in the string and the position of the deletion. Reliabilities of these tests also varied considerably. The best reliabilities were yielded by items in which the deletion occurred in the middle.

No significant differences were found between word-cloze scores in the three languages, although French and German letter-cloze scores proved on the whole to be lower than English letter-cloze scores. These findings might easily be attributed to peculiarities of the relatively small sample both of persons and of passages, but they also seem to be in accord with expectations based on information theory, on account of the larger effective alphabets of French and German when account is taken of diacritical marks.

In addition to these findings about the nature of the tests, the studies of adult bilinguals indicated that native speakers of a language show considerable variation in their ability to restore texts in their native language. Their ability to restore texts in a second language, in which they have near-native proficiency, is slightly (but significantly) poorer than their ability to restore texts in their native language. Further, the scores in the two languages were substantially correlated with each other. This suggested that the ability to restore texts is somewhat independent of competence in a language as ordinarily defined.

To further ascertain the characteristics of cloze tests several studies were conducted in English on native speakers. In a study in which the subjects were Harvard upper classmen and graduate students it was found that the measure of the difficulty of the passage is sensitive to the particular set of words deleted when passages 205 words in length containing 20 deletions were used. This finding is in accord with Taylor's (27) finding for passages of 175 words containing 35 deletions. In the present study the sensitivity of scores to the particular word deleted was demonstrated by comparing the adjusted means of passages in which every tenth word was deleted with the adjusted means of the same passages in which the deletions were moved over by one word. The adjustments were made by means of applying analysis of covariance to the data in which a third (control) passage was given to both the experimental and control group. It was also shown, in a similar experimental design, that subjects depend considerably upon contextual cues from the total passage. In this case the experimental manipulation consisted of dividing the passages into ten-word items (with the sixth word deleted) and scrambling the sequence of these items.

The data from the last-mentioned experiment made it possible to examine what kinds of items, from the point of view of syntactical structure, were most susceptible to the influence of paragraph cues. It was shown, within the limits of the sample of items available, that prepositions and adjectives (i.e., "function words") tended not to be affected by paragraph cues while nouns and verbs (i.e., "content-bearing words") were susceptible to influences from the paragraph. When the scores achieved on a test consisting 20 totally unrelated items, taken from the entire sample of items available, were compared with the scores obtained on continuous passages and on scrambled passages it was apparent that this diminution of contextual material further decreased the scores. The fact that paragraph context acts as an extraneous source of variation is further substantiated by the fact that reliability of scrambled passages is considerably higher than reliability of connected passages. However, it was determined that there was no "cumulative" effect of context; i.e., the degree to which context determined the responses to the last eight items of a paragraph was no greater than that for the first eight items.

Despite the fact that Taylor named the cloze test because it involves making "closure" in the sense of the word established by Gestalt psychologists, the ability to do cloze tests failed to show any correlation with Thurstone's (12) "Four Letter Word Test" which is reported to show a high loading on the "speed of closure" factor. Data collected in an independent factor analytic study by Weinfeld (32) using 9th, 10th, and 11th grade school children made it possible to show that the ability to do cloze tests in one's native language is related to reasoning, to verbal ability, to the ability to write good themes, and to expressive fluency.

Students from five public and private secondary schools in the Boston area were used to test the feasibility of cloze tests as measures of foreign language proficiency. Two 20-deletion word-cloze passages of appropriate difficulty and reliability, and two sets of letter-cloze materials consisting of 9- and 11-letter strings with the middle letter deleted, were administered in the respective language to 205 third, fourth, and fifth year students of French, and to 63 second and third year students of German. The time limits were sufficient to allow most students to attempt the last item of each of the 4 tests; the total testing time for the four tests including instructions was 36 minutes. The students appeared to accept the tests readily and were able to carry out the instructions. The tests were scored by allowing credit whenever the student restored the text exactly as it had stood in the original.

Collateral data, secured wherever possible, included scores on the Carroll-Sapon foreign language aptitude test, teachers' grades in the language courses, students' academic averages, intelligence test scores, and scores on College Board foreign language examinations.

On the two word-cloze tests combined, with a maximum possible score of 40, third-year French students had an average score of 14.5, and fourth-year French students had an average score of 16.1, as compared with adult native average performance of 27.4. On the two letter-cloze tests, with a maximum possible score of 100, third-year French students had average scores of 43.7, and fourth-year French students had average scores of 45.2, as compared with adult native average scores of 78.1. Thus, third-year French students do about 53% to 56% of adult performance, while fourth-year students do about 58% of adult performance, in both word-cloze and letter-cloze tests. The similarity of the percentages suggests that the cloze-procedure provides a rational scale along which performance can be measured in meaningful units.

Somewhat different results were obtained for the high school students of German, but it should be pointed out that these results are for 2nd and 3rd year students rather than for 3rd and 4th year students as in French. Second-year German students made average scores of 7.2 (24% of adult native performance) on the word-cloze test, and 46.7 (61% of adult native performance) on the letter-cloze test. Third-year German students made average scores of 10.2 (35% of adult performance) on the word-cloze test, and 47.8 (62% of adult performance) on the letter-cloze test; thus the constant proportionality observed in the case of the French results failed to occur.

Although there were wide individual differences in performance, there were no significant differences between the average performance of adjacent high-school year-groups on any of the tests. Either the instruction itself fails to cause significant gains from the 2nd to the 3rd or from the 3rd to the 4th years, or the present tests are insensitive to the gains. Nevertheless, high school students are significantly different from adult native speakers in their performance.

The reliabilities of the word-cloze tests in the high school samples were only moderate. For the 40-item French word-cloze test they ranged from .37 to .74 in various samples, and for the 40-item German word-cloze test they were .66 and .84 in two samples. Standard errors of measurement for these scores were fairly constant, ranging from 2.2 to 2.6. The magnitudes of the reliability coefficients were largely a function of different ranges of ability.

The reliabilities of 100-item letter-cloze tests ranged from .46 to .75 for French and from .80 to .88 for German. Standard errors of measurament were distinctly lower in German, being 4.8 to 5.1 as compared with 6.6 to 6.8 in French.

The reliability of cloze-procedure materials is such that in order to assure the achievement in typical secondary-school groups of a reliability coefficient of .85, say, it would be necessary to administer a word-cloze test of about 180 items (of 205-word paragraphs) with about an 80-minute time-limit; for letter-cloze, tests of 250 items with a 30-minute time-limit would be necessary.

The study provided abundant evidence that paragraph-length word-cloze tests and letter-cloze tests measure somewhat different aspects of second-language proficiency. In both French and German, correlations between word-cloze and letter-cloze tests were far from unity even after correction for attenuation. This was true, incidentally, both for high-school students and for adult native speakers.

The study involved numerous computations of correlations between cloze-test scores and various other measures of foreign language achievement, but it cannot be said that a completely satisfactory assessment of the validity of word-cloze and letter-cloze tests was made. This would require a criterion considerably more ultimate in its nature than any that were available. Favoring validity is the evidence that cloze tests differentiate significantly between learners and those who have achieved native mastery of the language. Also favoring satisfactory validity is the evidence that at least the word-cloze tests correlated quite substantially with teachers' grades; letter-cloze scores tend to correlate much lower with teachers' grades. It may be suggested that word-cloze tests make greater demands on the ability of the learner to select appropriate words and grammatical forms to fit a context, whereas the letter-cloze test demands chiefly a sensitivity to the orthographic customs of a language system and a knowledge of the proper spelling of foreign language words.

Word-cloze scores also correlated reasonably well ($r = .616$, $N = 76$) with CEEB French scores, whereas the letter-cloze scores correlated with these scores only to the extent of .354 and added nothing significant in a multiple correlation. Nevertheless, the corrections for attenuation suggest that neither the word-cloze nor the letter-cloze test measures the same kind of proficiency as the CEEB language examination. The results reported in Chapter 4 are strong reasons for believing that cloze procedure tests depend to a considerable extent upon cognitive ability variables which are completely extraneous to foreign language success. That is to say, even an individual who has good mastery of a foreign language may not be able to demonstrate this mastery on a cloze-procedure test if he lacks certain other intellectual qualities such as reasoning ability and ideational fluency.

This conclusion is further supported by the fact that cloze tests in foreign languages had substantial correlations with intelligence tests, in several high school samples. This was particularly true of letter-cloze tests, which had correlations of .44 to .61 in several groups, as compared to correlations of .26 to .61 for the word-cloze test.

The pattern of correlations between cloze tests and language aptitude test scores is similar to that between cloze tests and teachers' grades; that is, the correlations of language aptitude scores are higher with word-cloze scores than with letter-cloze scores.

Item analysis of the word-cloze tests administered in the high schools revealed that content-bearing form classes, such as nouns, adjectives and verbs are much more difficult to guess from context, while certain function-words like prepositions are easy to guess. This easily anticipated result is in accord with the findings of the item-study of the experiment in which the items within passages were scrambled. For purposes of test construction, however, it is important to note that both high and low item validities were found for both content-bearing form classes and for function words, so that little would be gained from systematically selecting one item type or another.

The conventional scoring system of counting as correct only the word given in the original text was compared with a community-of-response scoring system for French word-cloze data. The community-of-response scoring system yielded a higher reliability (using the passages as split halves of the test). The difference between the two reliabilities, however,. was not statistically significant. On the other hand the conventional scoring yielded a significantly higher correlation with CEEB scores than the community-of-response scoring system, thus suggesting that the conventional scoring system is more valid, or at least that the words used by the authors or translators are more in accord with the expectations of the writers of the CEEB tests than are the words used by the group whence the community of response was derived.

To explore the feasibility of testing in the auditory modality, two English auditory cloze tests were prepared on magnetic tape. The tests consisted of 28 items, each made up of six words, a deletion (signified by a standard noise introduced into the tape), and another four words. The consecutive items of each test constituted a passage. Each passage was prepared in an expressive form--using normal intonation and speech rate--and a non-expressive form--using a standard intonation pattern for each word, and a constant rate of presentation. A Latin square design was used so that each passage and each form was presented to all twenty English-speaking subjects. It was found that although the auditory cloze procedure proved to be a feasible measure of individual differences, as indicated by reliabilities yielded by the data and by the frequency distribution, the test form was highly unpleasant for the subjects. No difference was found between the mean scores on the expressive and non-expressive test forms. This finding suggests that auditory cloze scores will remain stable over a wide range of differences in intonation and rate of speakers of the material presented to the subjects. It does not imply that intonation plays no role in comprehension for it is probable that intonation carries information on a band of communication not measured by cloze procedure.

Conclusions

1. The word-cloze and letter-cloze procedures developed here may be suitable testing devices to assess group differences in second-language competence, but they are inadequate as measures of individual differences because they are relatively unreliable and are too heavily affected by various sources of extraneous variance. They require large amounts of

examining time and lend themselves well only to testing in the written language.

2. Word-cloze and letter-cloze procedures as developed here are extremely simple to prepare, except for the fact that care must be exercised to develop them from suitably chosen texts. Since they require free responses, however, they are somewhat cumbersome to score even though the scoring can be made completely objective.

Recommendations

1. It is not recommended that word-cloze and letter-cloze tests of the type investigated here be seriously considered by the CEEB as measures of foreign language achievement for use in Board examinations.

2. There are certain suggestions in the present study as to the lines along which further investigation might profitably proceed:

a. It may be suggested that the sources of extraneous variance in cloze procedure might be controlled either statistically--by adjusting for the individual's "ability to do cloze tests" as measured by a test or tests in his native language--or experimentally-- by searching for new types of cloze tests which are minimally dependent upon extraneous sources of variance.

b. Word-cloze tests consisting of relatively small segments of a text--say, ten-word items--may be a suitable way to reduce the influence of extraneous variance. Such tests would operate by minimizing the extent to which the examinee could utilize broad semantic cues afforded by paragraph-length context; the examinee would thus be forced to rely upon more purely linguistic cues such as the syntactical structure of the material.

c. The search for new types of cloze-procedure tests should concentrate on tasks in which native speakers of a language perform in a relatively uniform manner, but in which language learners progressively improve.

d. The use of a multiple-choice type of cloze procedure, in which the correct response is offered among a number of alternatives, deserves investigation. This type of item would have the advantage of possibly reducing the extent to which extraneous variance affects scores, while at the same time having the disadvantage of failing to test the examinee's ability to *produce* responses. It would also have the practical difficulty of requiring test construction efforts in the creation of alternatives and in standardizing scores.

e. If cloze-procedure is defined as any procedure in which the examinee is required to supply an element which will properly restore a mutilated text, it may be suggested that other kinds of cloze-procedure tests than the one investigated here might deserve investigation. For example, it might be found that Ebbinghaus's original procedure of requiring the subject to supply missing *syllables*, rather than letters or words, could be a useful technique for measuring foreign language proficiency without at the same time measuring various intellectual traits.

f. Besides cloze-procedure, there are undoubtedly many other ways of attempting to get at the individual's knowledge of the characteristics of a language structure without resorting to conventional test techniques. As examples of techniques which might deserve to be explored we may mention:

(1) Tests of the individual's ability to discriminate grammatical from ungrammatical sentences in the language (regardless of their degree of meaningfulness or nonsensicality), or to discriminate normal from mutilated texts.

(2) Tests of the individual's ability to judge the relative frequency of linguistic items in a language, as compared with the ability of native speakers to do so.

g. Perhaps the most effective way of studying the above possibilities would be to use a factor-analytic approach which would study the intercorrelations of a wide variety of test techniques in a suitably large sample of language learners or native speakers. The study should include measures of all four skills in language proficiency--listening, speaking, reading, and writing, as well as both conventional and novel kinds of testing techniques.

REFERENCES

1. Anderson, I. H. and Dearborn, W. F. *The psychology of teaching reading*. New York: Ronald Press Co., 1952.

2. Bruner, J. S. Going beyond the information given. In J. S. Bruner *et al*., *Contemporary approaches to cognition*. Cambridge, Mass.: Harvard Univ. Press, 1957. Pp. 41-69.

3. Burton, N. G., and Licklider, J. C. R. Long-range constraints in the statistical structure of printed English. *Amer. J. Psychol*., 1955, 68 650-653.

4. Carroll, J. B. A factor analysis of verbal abilities. *Psychometrika*, 1941, 6, 279-307.

5. Carroll, J. B. The factorial representation of mental ability and academic achievement. *Educ. psychol. Measmt*., 1943, 4, 307-332.

6. Carroll, J. B. and Sapon, S. M. *Modern language aptitude test*. New York: Psychological Corp., 1958.

7. Chapanis, A. The reconstruction of abbreviated printed messages. *J. exp. Psychol*., 1954, 48, 496-510.

8. Cofer, C. N. Cloze procedure in the evaluation of prose stories varying in adjective-verb quotient. Technical Report No. 13 for Contract NONR 595(04), Department of Psychology, U. of Maryland, College Park, Md. July, 1957.

9. Cofer, C. N. and Jenkins, Patricia M. A study of the learning of two stories by means of cloze technique. Technical Report No. 14 for Contract NONR 595(04), Department of Psychology, U. of Maryland, College Park, Md. July, 1957.

10. Davidson, W. M. and Carroll, J. B. Speed and level components in time-limit scores: a factor analysis. *Educ. psychol. Measmt*., 1945, 5, 411-427.

11. Ebbinghaus, H. Ueber eine neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bein Schulkindern. *Zeitschrift für Psychologie*, 1897, 13, 401-459.

12. Flesch, R. *Marks of readable style*. New York: Bureau of Publications, Columbia Univ., 1943.

13. French, J. W. (Ed.) *Manual for kit of selected tests for reference aptitude and achievement factors*. Princeton, N. J., Educational Testing Service, 1954.

14. Ebel, R. L. The estimation of the reliability of ratings. *Psychometrika*, 1951, 16, 407-424.

15. Howes, D. and Osgood, C.H. On the combination of associative probabilities in linguistic contexts. *Amer. J. Psychol.*, 1954, 67, 241-258.

16. Licklider, J.C.R. and Miller, G. A. The intelligibility of speech. In S. S. Stevens, (Ed.) *Handbook of experimental psychology*. New York: Wiley, 1951, Pp. 1040-1074.

17. McNemar, Q. *Psychological statistics*. New York: Wiley, 1957. 2nd Edit.

18. Miller, G. A. Human memory and the storage of information. *IRE Transactions on Information Theory*, 1956, Vol IT-2, No. 3, 129-137.

19. Miller, G. A., and Beebe-Center, J. G. Some psychological methods for evaluating the quality of translations. *Mechanical Translation*, 1956, 3, 73-80.

20. Miller, G. A., and Friedman, Elizabeth A. The reconstruction of mutilated English texts. *Information and Control*, 1957, 1, 38-55.

21. Miller, G. A., and Selfridge, J. A. Verbal context and the recall of meaningful material. *Amer. J. Psychol.*, 1950, 63, 176-185.

22. Peters, C. C., and Van Voorhis, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.

23. Shannon, C.E. Prediction and entropy of printed English. *Bell System Technical Journal*, 1951, 30, 50-64.

24. Shannon, C. E., and Weaver, W. *The mathematical theory of communication*. Urbana: Univ. of Illinois Press, 1949.

25. Taylor, C. W., Smith, W. R., Ghiselin, B., *et. al*. Identification of communication abilities in military situations. Unclassified report Proj. 7719; Task 17053, WADC-TR-58-92, Contract 18(600)-1211, Dept. of Psychology, U. of Utah, Utah, June 1958.

26. Taylor, W. L. "Cloze procedure": a new tool for measuring readability. *Journalism Quarterly*, 1953, 30, 415-433.

27. Taylor, W. L. Application of 'cloze' and entropy measures to the study of contextual constraint in samples of continuous prose. Unpublished doctoral dissertation, Univ. of Illinois, 1954. [Abstracted in *Dissertation Abstracts*, 1955, 15, 464-465.]

28. Taylor, W. L. Cloze procedure. *Agrisearch*, 1956, 2, 1-4.

29. Taylor, W. L. Recent developments in the use of "cloze procedure." Journalism Quarterly, 1956, 33, 42-48, 99.

30. Taylor, W. L. 'Cloze' readability scores as indices of individual differences in comprehension and aptitude. J. appl. Psychol., 1957, 41, 19-26.

31. Thibault, Paula. Implications of experience with College Board language tests. Georgetown Univ. Monog. Series in Languages and Linguistics, 1953, 4, 21-29.

32. Weinfeld, F. D. A factor analytic approach to the measurment of differential effects of training: an evaluation of four methods of teaching English composition. Unpublished doctoral dissertation in progress, Graduate School of Education, Harvard University, 1959.

33. Woodworth, R. S. Experimental psychology. New York: Holt, 1938.

APPENDIX A

INFORMATION SHEET

NAME ______________________ TEST BOOKLET NUMBER _____

Please describe below (1) when, (2) where and (3) under what circumstances you learned German. (4) Indicate whether German is your native language.

(1) ____

(2) ____

(3) ____

(4) ____

Please describe below (1) when, (2) where and (3) under what circumstances you learned English. (4) Indicate whether English is your native language.

(1) ____

(2) ____

(3) ____

(4) ____

Try to give an estimate below of your fluency in reading, writing and speaking German. Indicate whether you have read widely in all kinds of German writing; whethdr you have read a good deal but only in certain areas; whether you have hardly had a chance to read German, etc. We are particularly interested in how much experience you have had in reading and writing German.

Fluency:

Reading ____

Writing ____

Speaking ____

Extent of Reading

Try to give a similar estimate of your fluencies in English and compare them with your German fluencies. We are particularly interested in how much experience you have had in reading and writing English.

Fluency:

Reading ____

Writing ____

Speaking ____

Comparison of German and English Fluencies ____________________

____

Fill the blanks with a word which seems to fit best.

1. these creatures wear self- _________ coverings, to be sure of

2. If on closer examination _________ should then prove that these

3. regard these creatures as _________ despite their lack of language

4. earth, the pygmies. Let _________ suppose that in searching for

5. nothing can be _________ a language, either of sounds

6. and to consider them _________ be closely related to the

7. they do not employ _________ in their social intercourse. They

8. "pygmies" possess a fine _________ tail, which they use in

9. hitherto unknown primitive tribes _______ expedition stumbles unexpectedly upon some

10. would confirm the impression _________ they were human beings. We

11. we have attributed greater _________ to clothing than to speech.

12. make a thorough study _________ the most primitive people on

13. hypothetical speechless primitive men. _________ it then be discovered that

14. the most primitive sort, _________ protect themselves or, from feelings

15. pass over the fact _________ in classifying these problematical creatures

16. or of gestures. On _________ ground of their external similarity

17. of modesty, to cover _________ of their bodies, this discovery

18. to the pygmies, it _________ decided, with some misgivings, to

19. make emotional sounds somewhat _________ the apes; but they have

20. man-like creatures who differ _________ the pygmies only in that

Name: ______________________________

Age: _______

High School Class ________________ French Class ________________

How long have you studied French? __________ years

Have you ever studied any other foreign language? ____ yes ____ no (check one)

If 'yes', what other foreign language did you study? ____________________

For how many years? __________

Is a foreign language spoken in your home? ____ yes ____ no (check one)

INSTRUCTIONS FOR PART I

Part I of this test is made up of two French passages, each from a different book. Every tenth word has been replaced by a blank. Your task is to read through the passage to see what it is about, and then try to fill in each blank with a French word that makes sense. The blanks are all the same length, but remember that the words which have been taken out may be short or long. As some blanks will be easier to fill than others, it is a good idea to do the easier ones first, and then go back to work on the harder ones, if you have time.

Do one page at a time. When you finish one page, wait until your instructor tells you to go to the next. Do not look at the first page until your instructor tells you to begin.

Work rapidly, because you have only nine minutes for each passage!

CF 3

Le chef ne voulait pas être surpassé par les _________1, aussi descendit-il de cheval et prit-it également _________2 bâtons, mais Koumalo voyait bien qu'il ne comprenait _________3 tout à fait de quoi il s'agissait. Jarvis _________4 paraissait avoir la direction des opérations planta l'un _________5 bâtons dans le sol et le chef en passa _________6 à l'un de ses conseillers en lui disant _________7 chose. Alors le conseiller planta lui aussi son bâton _________8 le sol, mais l'homme à la boîte sur _________9 pieds cria:

--Pas là, pas là, enlevez-le. Sur _________10 le chef, embarrassé et sachant de moins en moins _________11 faire, remonta sur son cheval et y resta, laissant _________12 hommes blancs planter lerus bâtons.

Une heure s'écoula _________13 bout de laquelle tout un déploiement de bâtons et _________14 drapeaux se dressait, et Koumalo regardait toujours, de plus _________15 plus stupéfait. Jarvis et le magistrat parlaient ensemble et _________16 à se désigner tour à tour les collines et _________17 vallée. Puis ils s'adressèrent au chef, tandis que _________18 conseillers écoutaient gravement et attentivement leur conversation. Koumalo _________19 Jarvis dire au magistrat: --C'est trop long. Le _________20 haussa les épaules en disant:

. . . .

ç F 10

J'ai grimpé tant bien que mal sur un ________1 siège assez mal commode et presque aussitôt la longue ________2 à laquelle nous faisions face a paru bondir derrière ________3 tandis que la haute voix du moteur s'élevait ________4 cesse jusqu'à ne plus donner qu'une seule ________5, d'une extraordinaire pureté. Elle était comme le chant ________6 la lumière, elle était la lumière même, et je ________7 la suivre des yeux dans sa courbe immense, sa ________8 ascension. La Paysage ne venait pas à nous, il ________9 ouvrait de toutes parts, et un peu au-delà ________10 glissement hagard de la route, tournait majestueusement sur lui- ________11, ainsi que la porte d'un autre monde.

J' ________12 bien incapable de mesurer le chemin parcouru, ni le ________13. Je sais seulement que nous allions vite, très vite, ________14 plus en plus vite. Le vent de la course ________15 était plus, comme au début, l'obstacle auquel je ________16 appuyais de tout mon poids, il était devenu un ________17 vertigineux, un vide entre deux colonnes d'air brassées ________18 une vitesse foudroyante. Je les sentais rouler à ma ________19 et à ma gauche, pareilles à deux murailles liquides, ________20 lorsque j'essayais d'écarter . . . .

INSTRUCTIONS FOR PART II

On the pages that follow are words and parts of words. The first page has est items 11 letters long; the second one has examples 7 letters long. An item does not necessarily begin at the beginning of a word or end at the end of a word, and the examples are completely unrelated to each other. The middle letter of each example has been replaced by a blank.

Try to put a letter in the blank that makes sense. Besides the ordinary letters of the French alphabet, there are two other symbols you may use: the apostrophe, and "*", which stands for the space between words.

Sample problems	Answers
DEUX*_EURES	DEUX*HEURES
ENTIO_*DE*V	ENTION*DE*V
SON*F_ÈRE*E	SON*FRÈRE*E
PLU_*TA	PLUS*TA
*QU_ELL	*QU'ELL
CIR_CE*	CIR*CE*

Do not look at the next page until your instructor tells you to begin. When you finish one page, wait until you are told you may go on to the next.

FRANÇAIS

= F 11 b

SYMBOLS

A À Â B C C D E Ê É È F G H I J K L M N O Ô P Q R S T U Ù Û V W X Y Z ' * (Espace)

1. OURS*_'EXCE
2. QU'EL_E*Y*P
3. CTÈRI_TIQUE
4. ME*GU_RE*AV
5. AGE*D_HIVER
6. *CONT_NTES*
7. ÉRIVE_C'EST
8. C*LA*_LUS*G
9. ME*ET_QUE*S
10. C*IMP_TIENC
11. ELLE*_E*REG
12. *JE*N_*SAIS
13. CREUS_IT*PO
14. ORTAN_E*IL*
15. S*ET*_ILLES
16. L*N'E_T*PAS
17. SIQUE_ENTRE
18. T*IMP_RTANT
19. OPTER_DÉFIN
20. *MOIN_RE*PR
21. QUI*E_T*HUM
22. UN*J)_R*NOU
23. ANT*D_*RAME
24. COUP*_NE*TE
25. ENS*S_ÉCROU
26. UN*DE_I-SOM
27. TOUT*_U*MOI
28. ENT*P_R*QUA
29. TEUR*_'ARRÉ
30. *UN*A_TRE*M
31. *SE*R_CHAUF
32. NS*AP_ÈS*LE
33. *FLOT_ANT*D
34. NALEM_NT*DE
35. L'AUD_TOIRE
36. YANT*_N*N'A
37. CIEUS_S*ET*
38. IS*À*_A*REC
39. RS*JO_RS*OÙ
40. NE*SÉ_URITÉ
41. LLE*C_UT*QU
42. ND*JE_LUI*D
43. AIS*S_UVÉE*
44. À*CHA_TER*S
45. D*BAG_GE*À*
46. *ET*D_UNE*B
47. DONNA_LES*C
48. CARIB_U*COM
49. *ELLE_NE*SE
50. *LUI*_VAIT*

APPENDIX C (cont.)

FRANÇAIS

= F 9 b

SYMBOLS

A À Â B C Ç D E É È Ê F G H I J K L M N O Ô P Q R S T U Ù Û V W X Y Z ' , *
(Espace)

1. *QU_*TU
2. À*P_ÉSE
3. *LA_VIL
4. CAP_BLE
5. E*D_RE*
6. OUJ_URS
7. TOU_*LE
8. *CO_MEN
9. TRE_PIÈ
10. BIE_*MA
11. INQ_IÈT
12. INT_MIT
13. AVE_*UN
14. *NO_S*S
15. T*I_S*R
16. ET*_ANG
17. E*L_S*V
18. OME_T*M
19. AND_R*À
20. S*V_YON
21. ME*_HEZ
22. CI*_IT*
23. OUT_DE*
24. *DE_NOU
25. NTR_BÂI
26. AS*_IEN
27. VRE_MAR
28. ION_*PU
29. UR*_ETT
30. *MA_S*Ç
31. UT*_A*R
32. DE*_E*P
33. N*M_T*M
34. DIG_E*A
35. LA*_ORT
36. *NU_*À*
37. NCH_R*L
38. R*M_IS*
39. *TR_IS*
40. L'U_E*S
41. NGE_ET*
42. *PA_A*R
43. AN*_'EM
44. SAT_ON*
45. MAI_*À*
46. VIL_E*A
47. OMM_DE*
48. *TO_JOU
49. *À*_ABL
50. ARD_R*C

NAME: ______________________________

Age: __________

High School Class: ________________ German Class ____________________

How long have you studied German? ____________________ years

Have you ever studied any other foreign language? ___ yes ___ No (Check one)

If 'yes', what other foreign language did you study? ____________________

For how many years? ___________

Is a foreign language spoken in your home? ___ yes ___ No (Check one)

If 'yes', what language? _________________

INSTRUCTIONS FOR PART I

Part I of this test is made up of two German passages, each from a different book. Every tenth word has been replaced by a blank. Your task is to read through the passages to see what it is about, and then try to fill in each blank with a German word that makes sense. The blanks are all the same length, but remember that the words which have been taken out may be short or long. As some blanks will be easier to fill than others, it is a good idea to do the easier ones first, and then go back to work on the harder ones, if you have time.

Do one page at a time. When you finish one page, wait until your instructor tells you to go to the next. Do not look at the first page until your instructor tells you to begin.

Work rapidly, because you have only nine minutes for each passage!

C 30

Nun wollte sich der Häuptling von den weissen Männern ________ 1 ausstechen lassen, also stieg er auch vom Pferd und ________ 2 ein paar Stöcke, aber Kumalo sah wohl, dass er ________ 3 recht wusste, was eigentlich vorging. Jarvis, der die Sache ________ 4 leiten schien, steckte einen Stock in den Boden, und ________ 5 Häuptling gab einem seiner Räte einen Stock und sagte ________ 6 zu ihm. Also steckte dieser Mann auch den Stock ________ 7 den Boden, aber der weisse Mann mit dem Kasten ________ 8 drei Beinen rief:- Nicht dort, nicht dort; nimm den ________ 9 weg. Der Mann sah zweifelnd und zögernd den Häuptling ________ 10, und der sagte ärgerlich:- Nicht dort, nicht dort; nimm ________ 11 weg. Und dann stieg er wieder auf sein Pferd, ________ 12 und noch weniger als vorher begreifend, und sass da ________ 13 liess die weissen Männer ihre Stöcke in den Boden ________ 14.

So verging eine Stunde, während derer sich eine ganze ________ 15 von Stöcken un Fahnen bildete, und Kumalo sah zu ________ 16 wusste nach wie vor überhaupt nicht, was eigentlich vorging. Jarvis ________ 17 der Bürgermeister standen beieinander, und sie deuteten immer ________ 18 nach den Bergen hinauf, und dann wandten sie sich ________ 19 deuteten ins Tal hinunter. Dann sprachen sie mit dem ________ 20, und die Räte standen dabei . . . .

ç G 14

Widerwillig öffnete Jody die Augen. Manchmal überlegte er sich, _________ 1 schön es sein müsste, in den Wald zu entwischen _________ 2 dort ungestört von Freitag bis Montag zu schlafen. Durch _________ 3 Ostfenster seines kleinen Zimmers drang das Tageslicht herein. Er _________ 4 nicht recht, ob ihn die fahle Helle geweckt hatte _________ 5 das Gegacker der Hühner, die sich unter den Pfirsichbäumen _________ 6 schaffen machten. Er hörte, wie sie, eines nach dem _________ 7, von ihren Schlagplätzen im Baum zur Erde flatterten. Im _________ 8 färbte sich der Himmel mit gelbroten Streifen, vor denen _________ 9 die Umrisse der Fichten standen. Die Sonne ging im _________ 10 bereits merklich früher auf. Es konnte noch nicht sehr _________ 11 sein, und es war angenehm, aufzuwachen, ehe ihn die _________ 12 rief. Geniesserisch drehte er sich noch einmal auf die _________ 13 Seite. Das trockene Maisstroh seines Bettes raschelte unter ihm.

_________ 14 leuchtenden Streifen im Osten wurden heller. Ein goldener Strahl _________ 15 die Spitzen der Fichten, und während er noch hinsah, _________ 16 sich die Sonne selbst als glühender Ball. Wie von _________ 17 wachsenden Licht aus Osten herbeigeblasen rührte sich ein Lüftchen. _________ 18 Rupfenvorhänge wehten ins Zimmer. Nun strich die Brise über _________ 19 Bett und glitt weich, wie eine streichelnde Hand, über Jody _________ 20. Es war noch so warm . . . .

INSTRUCTIONS FOR PART II

On the pages that follow are words and parts of words. The first page has examples 11 letters long; the second one has examples 7 letters long. An example does not necessarily begin at the beginning of a word or end at the end of a word, and the examples are completely unrelated to each other.

Here the middle letter has been replaced by a blank: try to put a letter in the blank that makes sense. Besides the ordinary letters of the German alphabet, there is one other symbol you may use: "*", which stands for the space between words.

Sample Problems	Answers
JED*_OCHE*	JEDE*WOCHE*
ÄCHST_N*KAM	ÄCHSTEN*KAM
INMAL_SPOTT	INMAL*SPOTT
*ES_KLO	*ES*KLO
HM*_IT*	HM*MIT*
HEU_E*U	HEUTE*U

Do not look at the next page until your instructor tells you to begin. When you finish one page, wait until you are told you may go on to the next.

DEUTSCH

= G 11b

*MOGLICHE*ZEICHEN

* A Ä B C D E F G H I J K L M N O Ö P Q R S T U Ü V W X Y Z

1. UCH*F_HRT*E
2. N*MÜS_EN*SI
3. IND*D_S*SCH
4. R*JUN_E*KAP
5. N*HAN_ELSMA
6. *FLEH_NTLIC
7. ERKLÄ_TE*SI
8. EN*SE_HS*ST
9. RSTE*_ORGEN
10. *GRAU_E*GLI
11. *MEER_HINAU
12. VERSC_WINDE
13. DIE*M_NSCHE
14. GEHOL_EN*HA
15. AUF*I_MER*E
16. OCHE*_TAND*
17. WAR*I_R*DAS
18. UNDLI_HE*MI
19. GENOM_EN*UN
20. ICHT*_ASS*I
21. USCHT_WURDE
22. R*DIE_SCHMU
23. DUNKE_N*STR
24. ER*DE_*DIE*
25. CHT*V_RSANK
26. *BIS*_ER*LE
27. EL*VE_BLASS
28. *IN*D_R*NAC
29. DER*D_NKLE*
30. *TAG*_ETRAT
31. *DES*_ANDES
32. RÄUME_BEHER
33. HRE*I_RFAHR
34. MISSI_N*IN*
35. *AUFH_ELT*E
36. GEGEN_*DA*S
37. NGEN*_EISE*
38. N*IHR_EINEN
39. HRER*_IT*SI
40. DETE*_ND*VO
41. IGEN*_LTEN*
42. INEM*_ONAT*
43. ERNKL_IDER*
44. *LETZ_E*STR
45. U*BEW_LTIGE
46. CHMIT_AG*UM
47. ITTEN_WAREN
48. MAULT_ERE*Z
49. MIT*S_INEM*
50. R*INS_LAND*

= G 7b

DEUTSCH

MOGLICHE*ZEICHEN*

* A Ä B C D E F G H I J K L M N O Ö P Q R S T U Ü V W X Y Z

1. DES_VOL
2. LEN_*UN
3. LEG_N*S
4. UNT_RTA
5. DAS_URW
6. GES_ALT
7. UNT_R*D
8. SEI_*UN
9. UND_STE
10. FEN_ERZ
11. SS*_U*S
12. ESE_*EI
13. NE*_EST
14. SOL_TE*
15. CHE_*LA
16. R*D_R*F
17. IN*_ENS
18. ELL_SIG
19. ANN_NAC
20. *DA_*GL
21. ERP_ESS
22. TE*_IKT
23. R*I_N*K
24. ER*_EIN
25. ER*_IED
26. VOM_STA
27. SCH_S*F
28. BEI_PEI
29. LS*_R*A
30. SIC_ERT
31. ISC_E*V
32. *AL_EM*
33. GEN_SEI
34. BAL_TE*
35. CHE_*GL
36. BRU_ALE
37. ORT_ODO
38. IN*_ESE
39. ICH_*DE
40. ND*_CHO
41. BEN_GEB
42. WÄH_END
43. UM*_HN*
44. OPF_R*G
45. WEL_HES
46. STE_LUN
47. NWE_DUN
48. *EI_E*K
49. G*J_HRE
50. ZÄH_R*U

ANSWER KEYS

Appendix B, page 124. Sample Scrambled Test

1.	made	6.	to	11.	importance	16.	the
2.	it	7.	language	12.	of	17.	parts
3.	human	8.	long	13.	Should	18.	is
4.	us	9.	the	14.	to	19.	like
5.	called	10.	that	15.	that	20.	from

Appendix C, pages 126-127. French Word-Cloze Passages

ç F 3

1.	blancs	11.	que
2.	quelques	12.	les
3.	pas	13.	au
4.	qui	14.	de
5.	des	15.	en
6.	un	16.	continuaient
7.	quelque	17.	la
8.	dans	18.	les
9.	trois	19.	entendit
10.	quoi	20.	magistrat

ç F 10

1.	petit	11.	même
2.	descente	12.	étais
3.	nous	13.	temps
4.	sans	14.	de
5.	note	15.	n'
6.	de	16.	m'
7.	croyais	17.	couloir
8.	prodigieuse	18.	à
9.	s'	19.	droite
10.	du	20.	et

Appendix C, pages 129-130. French Letter-Cloze Items

= F 11b

1.	D	11.	M	21.	S	31.	É	41.	R
2.	L	12.	E	22.	U	32.	R	42.	*
3.	Ş	13.	A	23.	E	33.	T	43.	A
4.	E	14.	C	24.	U	34.	E	44.	N
5.	'	15.	F	25.	'	35.	I	45.	A
6.	E	16.	S	26.	M	36.	O	46.	D
7.	*	17.	*	27.	A	37.	E	47.	*
8.	P	18.	O	28.	A	38.	L	48.	O
9.	*	19.	*	29.	S	39.	U	49.	*
10.	A	20.	O	30.	U	40.	C	50.	A

= F 7b

1.	E	11.	U	21.	C	31.	L	41.	*
2.	R	12.	I	22.	D	32.	N	42.	P
3.	*	13.	C	23.	*	33.	O	43.	S
4.	A	14.	U	24.	*	34.	N	44.	I
5.	I	15.	L	25.	E	35.	P	45.	N
6.	O	16.	M	26.	B	36.	E	46.	L
7.	S	17.	E	27.	)	37.	E	47.	O
8.	M	18.	N	28.	S	38.	A	48.	U
9.	*	19.	E	29.	C	39.	O	49.	T
10.	N	20.	O	30.	I	40.	N	50.	E

Appendix D, pages 132-133. German Word-Cloze Passages

ç G 3

1.	nicht	11.	ihn
2.	nahm	12.	verwirrt
3.	nicht	13.	und
4.	zu	14.	stecken
5.	der	15.	Schlachtordnung
6.	etwas	16.	und
7.	in	17.	und
8.	auf	18.	wieder
9.	Stock	19.	und
10.	an	20.	Häuptling

ç G 14

1.	wie	11.	spät
2.	und	12.	Mutter
3.	das	13.	andere
4.	wusste	14.	die
5.	oder	15.	traf
6.	zu	16.	erhob
7.	anderen	17.	dem
8.	Osten	18.	die
9.	schwarz	19.	das
10.	April	20.	Jody or ihn

ANSWER KEYS
(continued)

Appendix D, pages 135-136. German Letter-Cloze Items

= G 11b

1. Ä	11. *	21. *	31. L	41. A
2. S	12. H	22. *	32. *	42. M
3. A	13. E	23. L	33. R	43. E
4. G	14. F	24. M	34. O	44. T
5. D	15. M	25. E	35. I	45. A
6. E	16. S	26. D	36. D	46. T
7. R	17. H	27. R	37. R	47. *
8. C	18. C	28. E	38. *	48. I
9. M	19. M	29. U	39. M	49. E
10. T	20. D	30. B	40. U	50. *

= G 7b

1. *	11. Z	21. R	31. H	41. *
2. D	12. N	22. D	32. L	42. R
3. E	13. B	23. H	33. *	43. E
4. E	14. L	24. M	34. L	44. I
5. *	15. N	25. B	35. M	45. E
6. T	16. M	26. *	36. T	46. C
7. E	17. O	27. E	37. H	47. L
8. N	18. *	28. S	38. D	48. N
9. *	19. S	29. E	39. T	49. N
10. H	20. R	30. H	40. S	50. A